Machine Learning-Based Hyperglycemia Prediction: Enhancing Risk Assessment in a Cohort of Undiagnosed Individuals

Kolapo Oyebola, Funmilayo Ligali, Afolabi Owoloye, Blessing Erinwusi, Yetunde Alo, Adesola Musa, Oluwagbemiga Aina, Babatunde Salako

Submitted to: JMIRx Med
on: February 01, 2024

Table of Contents

Machine Learning-Based Hyperglycemia Prediction: Enhancing Risk Assessment in a Cohort of Undiagnosed Individuals

Kolapo Oyebola[1, 2] DPhil; Funmilayo Ligali[1, 2] MS; Afolabi Owoloye[1, 2] MS; Blessing Erinwusi[2] MS; Yetunde Alo[2] MS; Adesola Musa[1] DPhil; Oluwagbemiga Aina[1] DPhil; Babatunde Salako[1] Dr med

[1]Nigerian Institute of Medical Research Lagos NG
[2]Centre for Genomic Research in Biomedicine, Mountain Top University Ibafo NG

Corresponding Author:
Kolapo Oyebola DPhil
Nigerian Institute of Medical Research
6, Edmund Crescent, Yaba
Lagos
NG

Abstract

Background: Noncommunicable diseases (NCDs) continue to pose a significant health challenge globally, with hyperglycemia serving as a prominent indicator of potential diabetes.

Objective: This study employed machine learning algorithms to predict hyperglycemia in a cohort of asymptomatic individuals and unraveled crucial predictors contributing to early risk identification.

Methods: This dataset included an extensive array of clinical and demographic data obtained from 195 asymptomatic adults residing in a suburban community in Nigeria. The study conducted a thorough comparison of multiple machine learning algorithms to ascertain the most effective model for predicting hyperglycemia. Moreover, we explored feature importance to pinpoint correlates of high blood glucose levels within the cohort.

Results: Elevated blood pressure and prehypertension were recorded in 8 (4%) and 18 (9%) individuals respectively. Forty-one (21%) individuals presented with hypertension (HTN), of which 34/41 (82.9%) were females. However, cohort-based gender adjustment showed that 34/118 (28.81%) females and 7/77 (9.02%) males were hypertensive. Age-based analysis revealed an inverse relationship between normotension and age ($r = -0.88$; $P < 0.05$). Conversely, HTN increased with age ($r = 0.53$; $P < 0.05$), peaking between 50-59 years. Isolated systolic hypertension (ISH) and isolated diastolic hypertension (IDH) were recorded in 16/195 (8.21%) and 15/195 (7.69%) individuals respectively, with females recording higher prevalence of ISH 11/16 (68.75%) while males reported a higher prevalence of IDH 11/15 (73.33%). Following class rebalancing, random forest classifier gave the best performance (Accuracy Score = 0.894; receiver operating characteristic-area under the curve (ROC-AUC) score = 0.893; F1 Score = 0.894) of the 27 model classifiers. The feature selection model identified uric acid and age as pivotal variables associated with hyperglycemia.

Conclusions: Random Forest classifier identified significant clinical correlates associated with hyperglycemia, offering valuable insights for early detection of diabetes and informing the design and deployment of therapeutic interventions. However, to achieve a more comprehensive understanding of each feature9s contribution to blood glucose levels, modeling additional relevant clinical features in larger datasets could be beneficial.

Keywords: Hyperglycemia; Diabetes; Machine Learning; Hypertension; Random Forest

(JMIR Preprints 01/02/2024:56993)
DOI: https://doi.org/10.2196/preprints.56993

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
Please make my preprint PDF available to anyone at any time (recommended).
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Machine Learning-Based Hyperglycemia Prediction: Enhancing Risk Assessment in a Cohort of Undiagnosed Individuals

Kolapo Oyebola[1,2,3], Funmilayo Ligali[1,2,3], Afolabi Owoloye[1,2,3], Blessing Erinwusi[1], Yetunde Alo[1], Adesola Musa[2], Oluwagbemiga Aina[2] and Babatunde Salako[2]

[1]Centre for Genomic Research in Biomedicine, Mountain Top University, Ibafo, Nigeria
[2]Nigerian Institute of Medical Research, Lagos, Nigeria
[3]Habilis Biotech Limited, Nigeria

Corresponding author: Kolapo Oyebola (PhD)

Centre for Genomic Research in Biomedicine, Mountain Top University, Ibafo, Nigeria
Email: oyebolakolapo@yahoo.com; kmoyebola@mtu.edu.ng
Phone: (+234)8034778549

Abstract

Background: Noncommunicable diseases (NCDs) continue to pose a significant health challenge globally, with hyperglycemia serving as a prominent indicator of diabetes.

Objective: This study employed machine learning algorithms to predict hyperglycemia in a cohort of asymptomatic individuals and unraveled crucial predictors contributing to early risk identification.

Methods: This dataset included an extensive array of clinical and demographic data obtained from 195 asymptomatic adults residing in a suburban community in Nigeria. The study conducted a thorough comparison of multiple machine learning algorithms to ascertain the most effective model for predicting hyperglycemia. Moreover, we explored feature importance to pinpoint correlates of high blood glucose levels within the cohort.

Results: Elevated blood pressure and prehypertension were recorded in eight (4%) and 18 (9%) of the 195 participants respectively. Forty-one (21%) participants presented with hypertension (HTN), of which 34/41 (83%) were females. However, gender adjustment showed that 34/118 (29%) females and 7/77 (9%) males were hypertensive. Age-based analysis revealed an inverse relationship between normotension and age (r = -0.88; P = .02). Conversely HTN increased with age (r = 0.53; P = .27), peaking between 50-59 years. Isolated systolic hypertension (ISH) and isolated diastolic hypertension (IDH) were recorded in 16/195 (8%) and 15/195 (8%) individuals respectively, with females recording higher prevalence of ISH 11/16 (69%) while males reported a higher prevalence of IDH 11/15 (73%). Following class rebalancing, Random Forest classifier gave the best performance (accuracy score = 0.89; receiver operating characteristic-area under the curve (ROC-AUC) score = 0.89; F1 Score = 0.89) of the 26 model classifiers. The feature selection model identified uric acid and age as important variables associated with hyperglycemia.

Conclusions: Random Forest classifier identified significant clinical correlates associated with hyperglycemia, offering valuable insights for early detection of diabetes and informing the design and deployment of therapeutic interventions. However, to achieve a more comprehensive understanding of each feature's contribution to blood glucose levels, modeling additional relevant clinical features in larger datasets could be beneficial.

Keywords: Hyperglycemia; Diabetes; Machine Learning; Hypertension; Random Forest

Introduction

Non-communicable diseases (NCDs) have become a significant public health concern in Africa [1]. Conditions like coronary artery disease, stroke, hypertension, and diabetes, which were once primarily associated with developed nations or affluence, have now become pervasive health challenges in developing countries and across diverse socio-economic strata [1]. The complex nature of NCDs underscores the need for a comprehensive approach to risk assessment, intervention and prevention.

Suburban communities serve as a distinctive microcosm within an evolving landscape of diseases [2, 3]. These communities, characterized by the coexistence of traditional and modern lifestyles, grapple with risk factors that necessitate thorough examination [4]. The epidemiological shift from communicable to non-communicable diseases, coupled with limited healthcare resources especially in suburban parts of developing countries [5, 6], stresses the importance of this research. In addition, recent advancements in genetic research have elucidated the underlying mechanisms of various complex NCDs. The identification of individuals at an elevated genetic risk for NCDs has the potential to revolutionize the approach of healthcare stakeholders to disease management. However, the effective implementation of genetic screening for NCD risk analysis relies on a robust understanding of the baseline contributors prevalent in the target population [7, 8]. This study provided a comprehensive description of the prevalence and intricate interplay of risk factors associated with NCDs, highlighting hypertension, obesity and diabetes. The specific focus was on undiagnosed asymptomatic individuals to elucidate the complex relationships of these health indicators within this population.

Machine learning encompasses a diverse set of algorithms designed to extract patterns from data and establish associations between these patterns and discrete sample classes within the data. Machine learning proves to be a valuable tool for identifying potential disease risk factors, elucidating etiology and interpreting complex pathological processes in the context of

NCDs [9-16]. In this study, multiple machine learning algorithms were developed to predict elevated blood glucose levels in a cohort of undiagnosed asymptomatic individuals. The primary objective was to systematically compare the accuracies of supervised machine learning classifiers to identify the most effective model for predicting hyperglycemia. Leveraging the predictors in the dataset, we meticulously constructed and evaluated these models for the identification of significant features associated with potential diabetes in the population.

Methods

Participant recruitment and screening

This study was carried out as part of a parallel community-based genetic screening of apparently healthy adults living in Ijede Community, Lagos, Nigeria. Ethical approval was obtained from the Institutional Review Board of the Nigerian Institute of Medical Research (IRB/21/074). Following informed consent, participants were recruited and 10ml of venous blood samples were collected per individual. Demographic information, body mass index (BMI), knowledge, attitude and practices were obtained from the participants. The study clinician further clerked participants for personal and family medical history as well as their smoking status. Exclusion criteria included pregnancy at the time of recruitment, placement on antihypertensive or antidiabetic chemotherapy, radiotherapy, current or previous hematologic or tumoral diseases and known chronic diseases. Participants underwent electrocardiogram (ECG) screening (SonoHealth, USA) to provide clues on heart defects or other heart-related problems. Hemoglobin electrophoresis was conducted to detect possible hemoglobinopathy in the participants [17]. In addition, random blood glucose concentrations (Guilin Royalze, China) and blood pressure (BP) values (Iston Mediq, USA) were determined to evaluate the presence or absence of prediabetes, diabetes, prehypertension (preHTN) or hypertension (HTN) onset in the participants. Individuals with screening tests outside normal ranges were advised to visit their healthcare specialists for further checks. Normal BP was described as systolic blood pressure (SBP) <120mmHg and diastolic blood pressure (DBP) <80

mmHg. Elevated BP was defined as SBP of 120–129 mmHg and DBP <80 mmHg, stage 1 hypertension (preHTN) as SBP ≥ 130-139 mmHg and DBP 80 – 89 mmHg and stage 2 HTN as SBP ≥140 and DBP ≥ 90 mmHg [18]. Isolated systolic hypertension (ISH) was described as SBP above 140 mmHg with diastolic blood pressure (DBP) of less than 90 mmHg [19]. Isolated diastolic hypertension (IDH) is an important subtype of hypertension defined as a systolic blood pressure (SBP) of <130 mm Hg and a diastolic blood pressure (DBP) of at least 80 mm Hg [20]. Prediabetes was defined as random blood glucose (RBG) concentration of 140–199 mg/dl or fasting blood glucose of 100–125 mg/dl. Diabetes mellitus was defined as random blood glucose level of ≥200 mg/dl or fasting blood glucose of ≥126 mg/dl [21]. However, as all the participants reported they were not fasting, random blood glucose values were documented.

Correlation analysis

Data cleaning, exploratory analysis and feature engineering were performed in Google Colab (with Python 3.10). The target variable was specified as "blood glucose," where 1 indicated a RBG concentration ≥140mg/dl and 0 indicated RBG concentration <140mg/dl. Independent variables included age (integer), sex (integer), BMI (float), smoking status (integer), ECG (float), hemoglobin (float), cholesterol (float), uric acid (float), systolic blood pressure (integer), diastolic blood pressure (integer), normal BP (integer), elevated BP (integer), preHTN (integer), HTN (integer), isolated systolic hypertension (integer), isolated diastolic hypertension (integer), prediabetes (integer), diabetes (integer), normal glucose (integer), abnormal ECG values (integer) and normal ECG values (integer). The dataset was checked and visualized for missingness using seaborn heatmap (Additional File 1: Fig. S1). Missing values were replaced with column mean (for continuous variables) or mode (for categorical variables). Duplicate rows and outliers were dropped before encoding categorical variables and creating dummy variables. Subsequently, we created a heatmap of correlation of independent variables with target column in descending order. The cleaned dataset was then scaled for subsequent training of machine learning models. *P*-value ≤ .05 was considered statistically significant.

Machine learning algorithms and evaluation

The study adopted 26 supervised classification algorithms and compared their accuracies to identify the best performing model for predicting high blood glucose which was defined in this study as

random blood glucose (RBG) concentration ≥140mg/dl (Fig. 1). Specifically, after installation and importation of Sci-Kit Learn libraries [22], we carried out data cleaning, exploration and scaling to improve the efficiency of our model (Supplementary Methods). Imbalances in the distribution of hyperglycemia cases and non-cases within the dataset might affect the model's performance. Addressing this imbalance and validating the model on balanced datasets could enhance its robustness. To address class imbalance in the outcome variable (blood glucose level), we adopted synthetic minority over-sampling technique (SMOTE). SMOTE tackled the underrepresentation of the minority class and rebalanced the class distribution for equitability [23]. After resampling, we split the data into training and test sets at ratio 80:20 respectively, using the train_test_split function in Sci-Kit Learn. We went further to select and rank the performances of the machine learning algorithms using LazyPredict to obtain the weighted average of the F1 and accuracy scores as well as the receiver operating characteristic-area under the curve (ROC-AUC) score. For hyperparameter optimization, we adopted GridSearchCV (https://github.com/oyebolakolapo/Machine-Learning-Prediction-of-Elevated-Blood-Glucose-in-a-Cohort-of-Apparently-Healthy-Adults). The grid search technique constructs many versions of the model with all possible combinations of hyperparameters to return the best one [24]. Subsequently, we determined feature importance to provide insight into which features are most associated with elevated blood glucose level using the best performing model. To operationalize the best performing model generated at scale, the training file was stored as a serialized pickle file. Subsequently, we used Fast application programming interface (Fast API) in Google Colab [25], to make an inference call from the model using the predict() function and generated our API. Pyngrok was used to open secure tunnels from public uniform resource locator (URL) to local host.

Results

Cohort description

Two hundred participants aged 18-83 years were enrolled into the cohort. However, after hemoglobin electrophoresis screening, five individuals were found to possess the HbSS/HbSC genotypes and were excluded from further analysis. Enlisted individuals consisted of 118 females and 77 males (Fig. 2; Additional

File 1: Fig. S2).

Correlation analysis

Participants were categorized into six age groups: 18-29; 30-39; 30-49; 50-59; 60-69 and =>70 years. Elevated blood pressure and preHTN were recorded in eight (4%) and 18 (9%) individuals respectively (Fig. 3). Forty-one (21%) of the cohort (n = 195) presented with HTN, of which 34/41 (83%) were females (Additional File 1: Fig. S3). Age-based analysis revealed an inverse relationship between normotension and age (r = -0.88; P = .02). Consistently, HTN increased with age (r = 0.53; P = .27), peaking between 50-59 years (Fig. 4). ISH and IDH were recorded in 16/195 (8%) and 15/195 (8%) participants respectively, with females recording higher prevalence of ISH 11/16 (69%) while males reported a higher prevalence of IDH 11/15 (73%) (Additional File 1: Fig. S4). There was a positive correlation between ISH and participants' age (r = 0.86; P = .03), whereas IDH was inversely correlated with age (r = -0.71; P = .11) (Fig. 5). We went further to examine the heart rates of the participants and observed an age-dependent increase in the percentage of participants with abnormal ECG values peaking between 60-69 years (Fig. 6). However, no significant difference was observed in the ECG values of male and female participants ($\sqrt{X^2}$ = 0.13, P = .72) (Additional File 1: Fig. S5). Random blood glucose value between 140 - 199mg/dl (prediabetes) was detected in 22 (12%) while diabetes was suspected in five (3%) of the 195 participants respectively (Fig. 7). Though not statistically significant, an inverse relationship (r = -0.81; P = .06) was observed between age and normal glucose level andthe frequency of prediabetes (r = 0.63; P = .19) and suspected diabetes (r = 0.58; P = .24) seemed to increase with age (Additional File 1: Fig. S6). Meanwhile, a correlation matrix between each independent variable and the target column (blood glucose level) showed that age had the highest ranking even though the correlation coefficient was weak (Fig. 8; Additional File 1: Fig. S7).

Machine learning algorithms and evaluation

Following data cleaning, transformation (Additional File 1: Fig. S8) and observation of a class imbalance in the target variable (Additional File 1: Fig. S9), whereby the raw dataset demonstrated that 163/195 (83.6%) of the participants had normal blood glucose {0} while 32/195 (16%) had high blood glucose level {1}, rebalancing was established with SMOTE to yield an even representation of

both categories of blood glucose level (Counter ({0: 163, 1: 163}). When the performance of each classifier was tested, the reports showed Random Forest Classifier (Figs. 9 and 10) gave the best accuracy (Accuracy Score = 0.89; ROC-AUC score = 0.89; F1 Score = 0.89) followed by Extra Trees (Accuracy Score = 0.88; ROC-AUC score = 0.88; F1 Score = 0.88) and XGB classifiers (Accuracy Score = 0.86; ROC-AUC score = 0.86; F1 Score = 0.86), respectively (Fig. 9B; Table 1).

To determine the importance of each variable (feature) to the outcome (blood glucose level), we carried out random forest feature analysis. The importance of a feature is calculated based on how much the tree nodes that use that feature reduce impurity across all trees in the forest. The key findings showed that uric acid and age were the most important features associated with elevated blood glucose (Fig. 11), followed by systolic blood pressure and body mass index (BMI).

Table 1: Performance of model classifiers following SMOTE rebalancing.

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken (s)
Random Forest Classifier	0.89	0.89	0.89	0.89	0.18
Extra Trees Classifier	0.88	0.87	0.87	0.88	0.15
XGB Classifier	0.86	0.86	0.86	0.86	0.06
NuSVC	0.86	0.86	0.86	0.86	0.02
LGBM Classifier	0.85	0.85	0.85	0.85	0.10
Label Propagation	0.85	0.84	0.84	0.85	0.02
Label Spreading	0.85	0.84	0.84	0.85	0.03
Bagging Classifier	0.83	0.83	0.83	0.83	0.04
K neighbors Classifier	0.83	0.83	0.83	0.83	0.02

Decision Tree Classifier	0.80	0.80	0.80	0.80	0.01
AdaBoost Classifier	0.76	0.75	0.75	0.76	0.13
SVC	0.73	0.73	0.73	0.73	0.02
SGD Classifier	0.70	0.69	0.69	0.69	0.01
Ridge Classifier CV	0.70	0.70	0.70	0.70	0.02
Linear Discriminant Analysis	0.68	0.69	0.69	0.68	0.02
Ridge Classifier	0.68	0.69	0.69	0.68	0.02
Linear SVC	0.68	0.68	0.68	0.68	0.05
Calibrated Classifier CV	0.67	0.67	0.67	0.67	0.09
Quadratic Discriminant Analysis	0.65	0.66	0.66	0.65	0.02
Logistic Regression	0.62	0.62	0.62	0.62	0.03
Bernoulli NB	0.62	0.62	0.62	0.62	0.01
Perceptron	0.59	0.60	0.60	0.59	0.01
Gaussian NB	0.53	0.54	0.54	0.52	0.02
Passive Aggressive Classifier	0.52	0.52	0.52	0.50	0.02
Nearest Centroid	0.52	0.52	0.52	0.51	0.01
Dummy Classifier	0.47	0.50	0.50	0.30	0.01

Discussion

Noncommunicable diseases, such as cancer, cardiovascular diseases, and diabetes, are progressively becoming the primary causes of mortality in sub-Saharan Africa [26]. This epidemiological shift is primarily attributed to limitations in implementing crucial control measures, such as prevention and early detection [1]. This research focused on exploring key clinical indices of NCDs in asymptomatic individuals. The application of machine learning in disease prediction is now well-established for its immense potential in analyzing complex datasets and uncovering patterns that may elude human detection [27-30]. The investigation employed various machine learning algorithms to predict hyperglycemia to enable early identification of individuals at a particular risk of developing diabetes. The study identified suspected hypertension in 21% of study participants, underscoring the urgency of addressing hypertension as a major health challenge in the country. Furthermore, a notable increase in the prevalence of hypertension with advancing age was observed. However, the investigation into hypertension subtypes revealed a dual phenomenon: a pronounced increase in systolic hypertension with age and a concomitant reduction in diastolic hypertension.

Several factors may contribute to the observed age-related increase in systolic hypertension. Physiological changes, alterations in vascular reactivity, and lifestyle factors could play decisive roles in driving the upward trajectory of systolic blood pressure with advancing age [31, 32]. In contrast, the age-related reduction in diastolic hypertension may be associated with changes in arterial compliance, heart rate dynamics, or other physiological adaptations over the aging process [33]. Recognizing these dual dynamics holds significant clinical implications, necessitating tailored screening protocols and interventions to address the unique challenges posed by hypertension in different age groups.

Moreover, a gender disparity was observed, with systolic hypertension being more prevalent in females while diastolic hypertension was more common in males. This gender difference may be linked to heart rate variability or hormonal influences, particularly fluctuations in estrogen levels in females. However, understanding how blood vessels respond to changes in pressure and the potential impact on systolic blood pressure would be crucial in deciphering these gender disparities [34-36]. Therefore, tailoring screening protocols and interventions to address the unique challenges posed by hypertension in different age groups and genders is essential to mitigate the overall burden of this condition.

Electrocardiography is a pivotal tool for assessing cardiac health, and its interpretation can provide valuable insights into cardiovascular conditions. Our investigation revealed a remarkable age-dependent pattern in abnormal ECG values, reaching a peak at 70 years. Advancing age often coincides with a myriad of physiological changes, including alterations in cardiac structure and function [37-39]. A comprehensive exploration of these factors is essential for delineating the intricate relationship between aging and abnormal ECG findings.

The global burden of diabetes is well-documented [40-42], but our investigation into supposedly healthy individuals has unearthed a concerning revelation. Despite outward appearances of health,

there existed a relatively high prevalence of suspected prediabetes and diabetes in the cohort. This underscores the importance of probing beyond outward health markers to understand latent metabolic landscape [43-46]. This prompts a reevaluation of health screening protocols to incorporate metabolic parameters in apparently healthy populations. Early detection and intervention strategies should be tailored to encompass metabolic assessments, providing an opportunity for targeted preventive measures and lifestyle modifications.

In the realm of predictive modeling, selecting the most effective machine learning algorithm is paramount. Our study, aimed at evaluating various algorithms, revealed insightful findings regarding their predictive performances. Upon meticulous evaluation, Random Forest emerged as the top-performing algorithm, consistently delivering the highest accuracy among the tested models. The success of the Random Forest algorithm can be attributed to its ensemble learning nature [47, 48], which harnesses the collective power of multiple decision trees. This enables robustness against overfitting, enhanced generalization, and effective handling of complex datasets with diverse features. The observed superiority of Random Forest in our study has profound implications for future applications, suggesting its applicability across diverse datasets and underscoring its potential as a reliable choice for achieving high predictive accuracy.

To investigate the intricate determinants of hyperglycemia, our study employed a robust feature importance analysis, with compelling results showcasing uric acid and age as the most influential predictors. Uric acid's prominence as a predictor of hyperglycemia adds a unique dimension to our understanding of metabolic health. While traditionally associated with conditions like gout, our findings suggest a potential link between hyperuricemia and hyperglycemia, urging further exploration into the underlying physiological mechanisms. The identification of age as a key predictor aligns with existing knowledge regarding the age-associated risk of hyperglycemia [48-50]. Our findings reinforce the significance of age as a robust indicator, reflecting the cumulative impact of aging processes on metabolic health and glucose regulation. The recognition of uric acid and age

as pivotal predictors holds significant clinical implications. Healthcare practitioners can leverage these findings to enhance risk assessment strategies for hyperglycemia. Incorporating uric acid measurements and age considerations into routine screenings may facilitate early identification of individuals at heightened risk, enabling proactive interventions. While our study sheds light on the importance of uric acid and age, further research is warranted to unravel the intricate relationships and mechanisms underlying these associations. Longitudinal studies exploring the dynamic interplay between uric acid, age and hyperglycemia can deepen our understanding and inform targeted interventions.

Limitations and Future Direction

While our study provides valuable insights into predicting hyperglycemia using machine learning in undiagnosed individuals, it is essential to acknowledge certain limitations that may impact interpretation. First, the size of our cohort may limit the generalizability of the results. A larger and more diverse sample could enhance external validity of the predictive model. Furthermore, the study did not account for potential variations in clinical practice, including differences in diagnostic criteria. For instance, the study did not take into consideration orthostatic hypotension, a fall in SBP of at least 20 mm Hg or a DBP fall of at least 10 mm Hg within three minutes of standing, especially in older individuals [19]. Although seats were provided to participants, we could not accurately document how long participants had been standing before attending the screening. Besides, phenomena such as postprandial hypotension (a reduction in BP after meals, a common cause of syncope and falls in healthy and hypertensive elderly individuals), circadian BP variability, and white-coat (non-sustained) hypertension, especially in the elderly were not factored into the analyses [51-53]. As such, incorporating standardized criteria across diverse healthcare settings could enhance our model's clinical applicability.

Moreover, the study did not dissect the influence of ethnicity and genetics on hyperglycemia [54, 55]. Future research could explore these aspects to provide a more comprehensive understanding of predictive factors. Since the dataset primarily comprises information from a specific geographic

location or demographic group, extrapolating the findings to other populations requires caution as regional variations in lifestyle, genetics, and healthcare practices may influence the performance of the predictive model. In addition, the cross-sectional nature of our study limits our ability to establish causation or assess changes over time. Therefore, longitudinal studies would be beneficial to understand the dynamic nature of hyperglycemia predictors. The model's performance was evaluated on the same dataset used for training, raising the potential for overfitting. External validation on an independent dataset would be crucial to assess its generalizability and reliability in real-world scenarios. Lastly, the importance of a feature in a Random Forest model does not necessarily mean a causal relationship and other models might find different results if additional features are introduced. Future approaches are expected to accommodate more features and larger datasets. This will account for the deployment of built and containerized models as publicly accessible web apps. Nevertheless, this present study has expounded the potential of machine learning for early disease detection, risk assessment strategies, proactive interventions and targeted therapeutic design.

Conclusions

This study has made a substantial contribution to the expanding domain of predictive modeling and offers promising implications for enhancing early detection and personalized risk assessment, particularly in the context of hyperglycemia and its potential association with diabetes. The research has not only brought to light the prevalence of undiagnosed hypertension, isolated systolic and diastolic hypertension but has also highlighted factors associated with elevated blood glucose within the population. The findings of this study emphasize the significance of regular screening, effective intervention strategies and targeted therapeutic designs. Collectively, the results contribute to the overarching effort to enhance healthcare outcomes through proactive and tailored approaches.

List of Abbreviations

API: Application Programming Interface

BMI: Body Mass Index

BP: Blood Pressure

DBP: Diastolic Blood Pressure

ECG: Electrocardiogram

HTN: Hypertension

IDH: Isolated Diastolic Hypertension

ISH: Isolated Systolic Hypertension

NCDs: Noncommunicable Diseases

RBG: Random Blood Glucose

ROC-AUC: Receiver Operating Characteristic-Area under the Curve

SBP: Systolic Blood Pressure

SMOTE: Synthetic Minority Over-Sampling Technique

URL: Uniform Resource Locator

Declarations

Ethics approval and consent to participate.

Ethical approval was obtained from the Institutional Review Board of the Nigerian Institute of Medical Research (IRB/21/074) in accordance with the Declaration of Helsinki [56]. Before enrollment, all the participants signed Informed consent forms.

Consent for publication

Not applicable.

Availability of Data and Materials

The dataset(s) supporting the conclusions of this article are within the manuscript and its Additional File 1 supplementary information.The Google Colab Python script used for data analysis and machine learning has been deposited in our GitHub page https://github.com/oyebolakolapo/Machine-Learning-Prediction-of-Elevated-Blood-Glucose-in-a-Cohort-of-Apparently-Healthy-Adults.

Competing Interests

Authors declare no competing interests.

Funding

Kolapo Oyebola was supported by Fogarty Emerging Global Leader Grant (NIH-K43TW011926)

from the US National Institutes of Health and an APTI-18-07 Grant from the African Academy of Sciences in partnership with Bill and Melinda Gates Foundation. The funders had no role in study design, data collection, analysis, decision to publish or preparation of the manuscript.

Authors' Contributions

KO conceived and designed the study. KO, FL, AO, BE, YA and OA implemented field study. KO, FL and BE carried out laboratory experiments. KO carried out data analysis, including machine learning. KO drafted the manuscript. KO, OA and BS edited the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors appreciate the study participants and Ijede Community Leaders for their cooperation during the screening exercise.

Authors' Information

Centre for Genomic Research in Biomedicine, Mountain Top University, Ibafo, Nigeria

Kolapo Oyebola, Funmilayo Ligali, Afolabi Owoloye, Blessing Erinwusi, Yetunde Alo

Nigerian Institute of Medical Research, Lagos, Nigeria

Kolapo Oyebola, Funmilayo Ligali, Afolabi Owoloye, Oluwagbemiga Aina and Babatunde Salako

Habilis Biotech Limited, Nigeria

Kolapo Oyebola, Funmilayo Ligali, Afolabi Owoloye

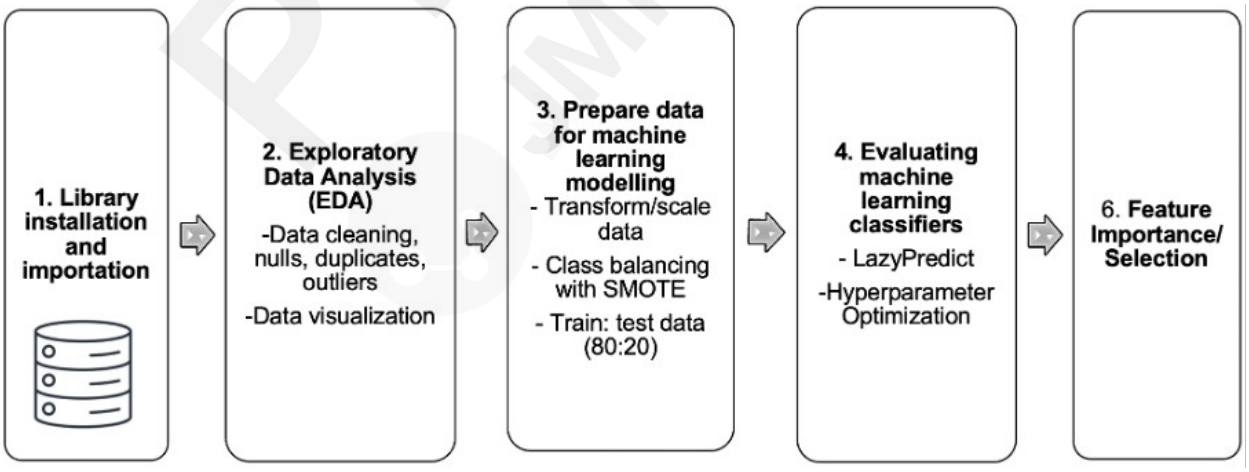


Figure 1: **Pipeline for model development**

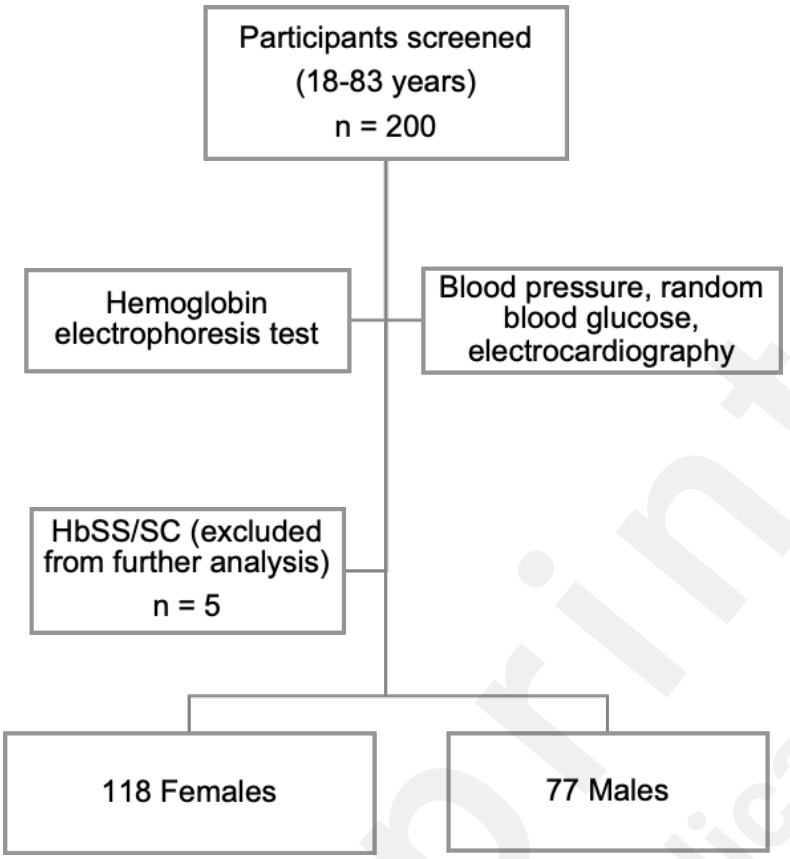


Figure 2: **Participant recruitment and screening**

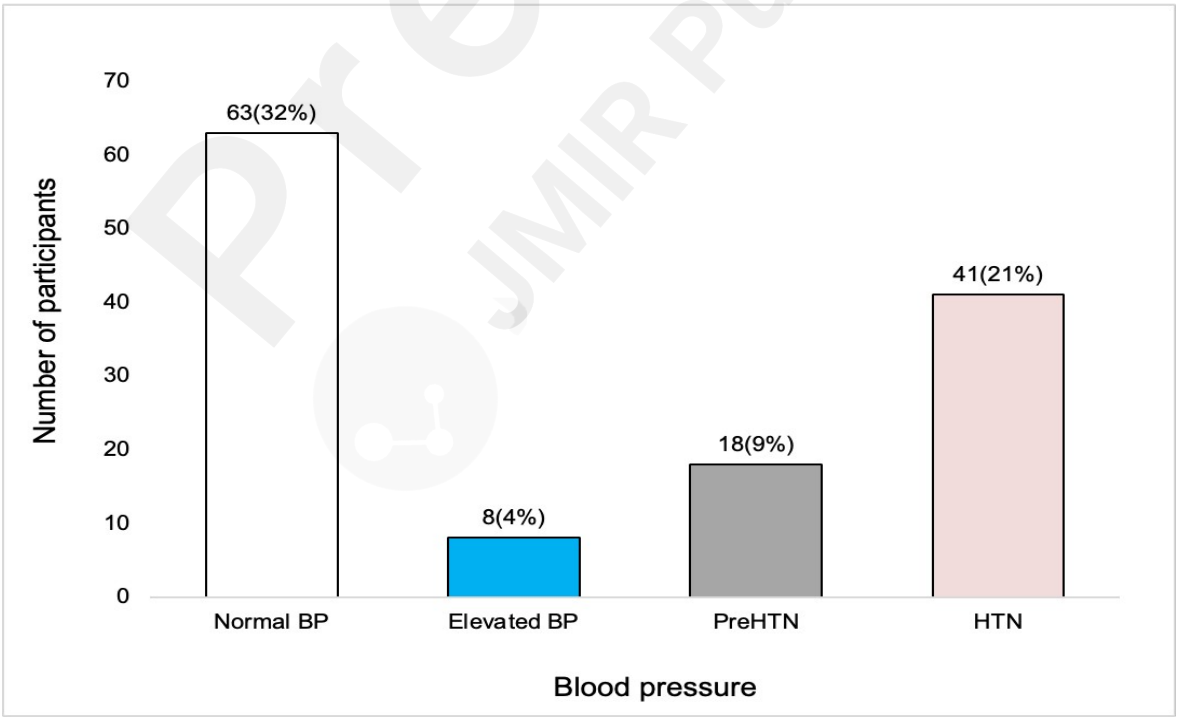


Figure 3: Blood pressure readings (BP = Blood Pressure; HTN = Hypertension)

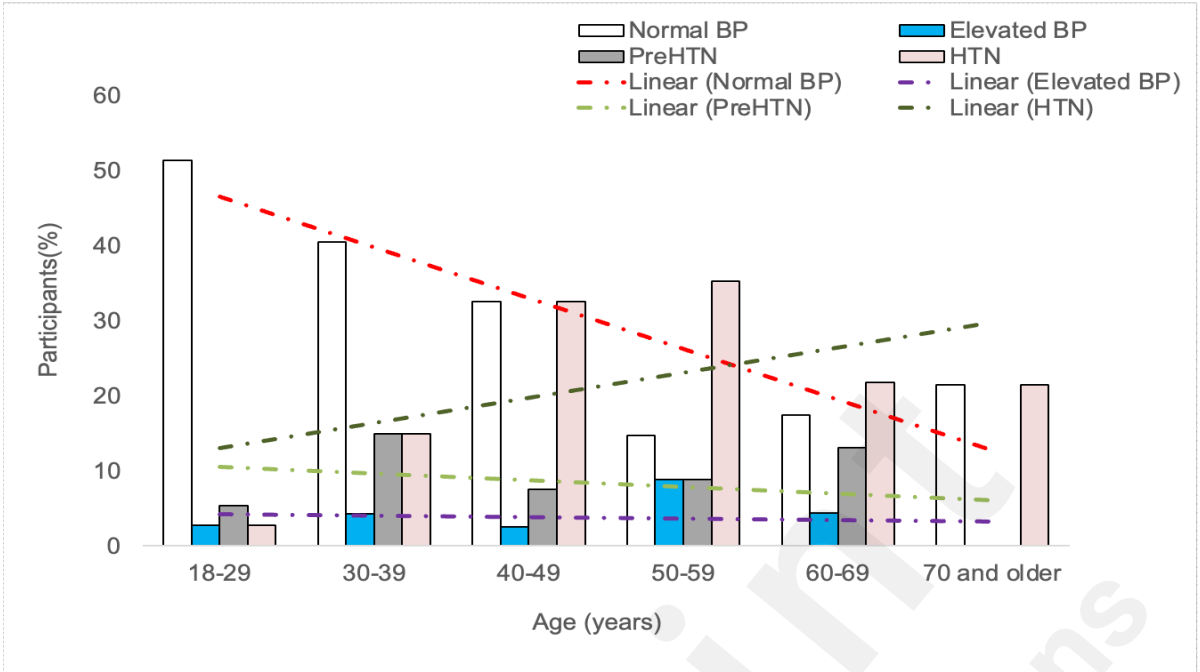


Figure 4: **Age-based analysis of blood pressure.** Percentage of participants with normal blood pressure (BP) reduced with increase in age (r = -0.88; P = .02). Prevalence of hypertension (HTN) increased with age (r = 0.53; P = .27), peaking between 50-59 years

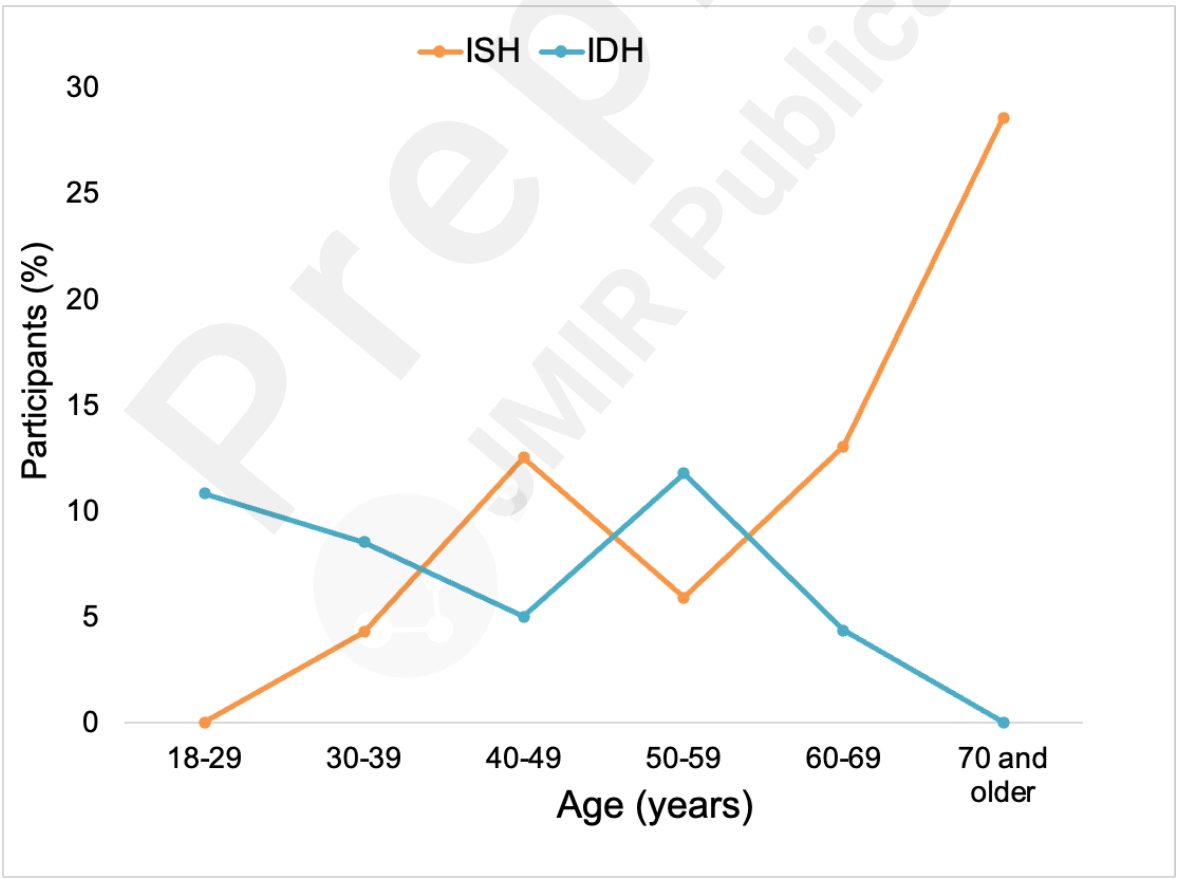


Figure 5: Age-based analysis of isolated systolic hypertension (ISH) and isolated diastolic hypertension (IDH). ISH increased with participants' age (r = 0.86; P = .03), unlike IDH (r = -0.71; P = .11)

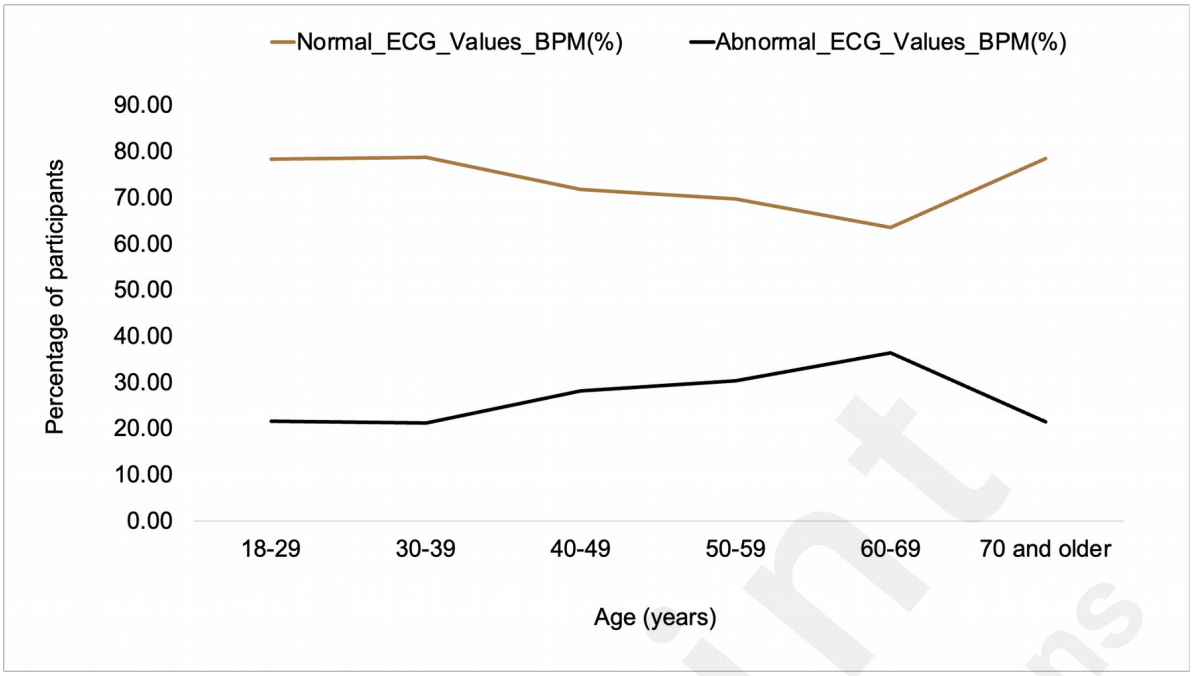


Figure 6: Age-based electrocardiogram (ECG) analysis. Age-dependent increase in the percentage of participants with abnormal ECG values peaking between 60-69 years.

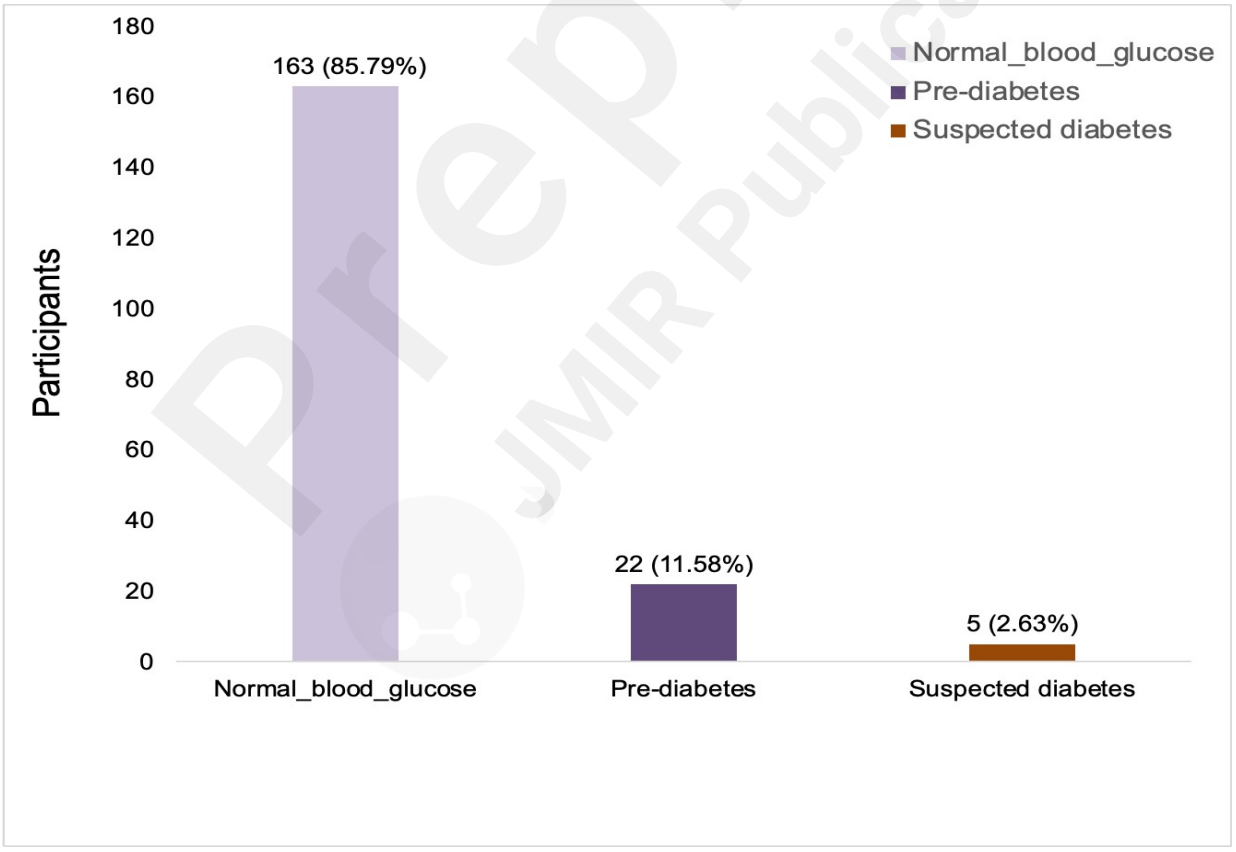


Figure 7: **Blood glucose levels in the cohort**

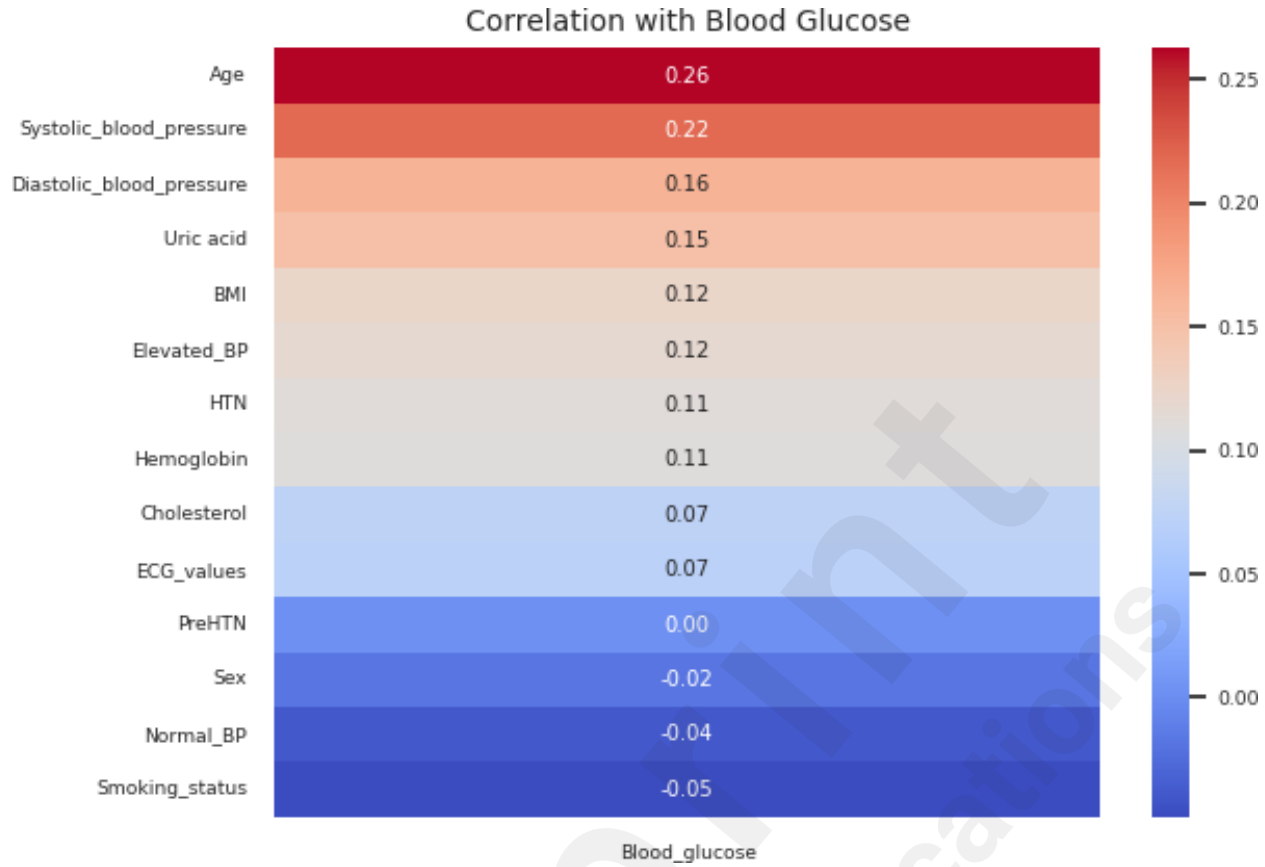
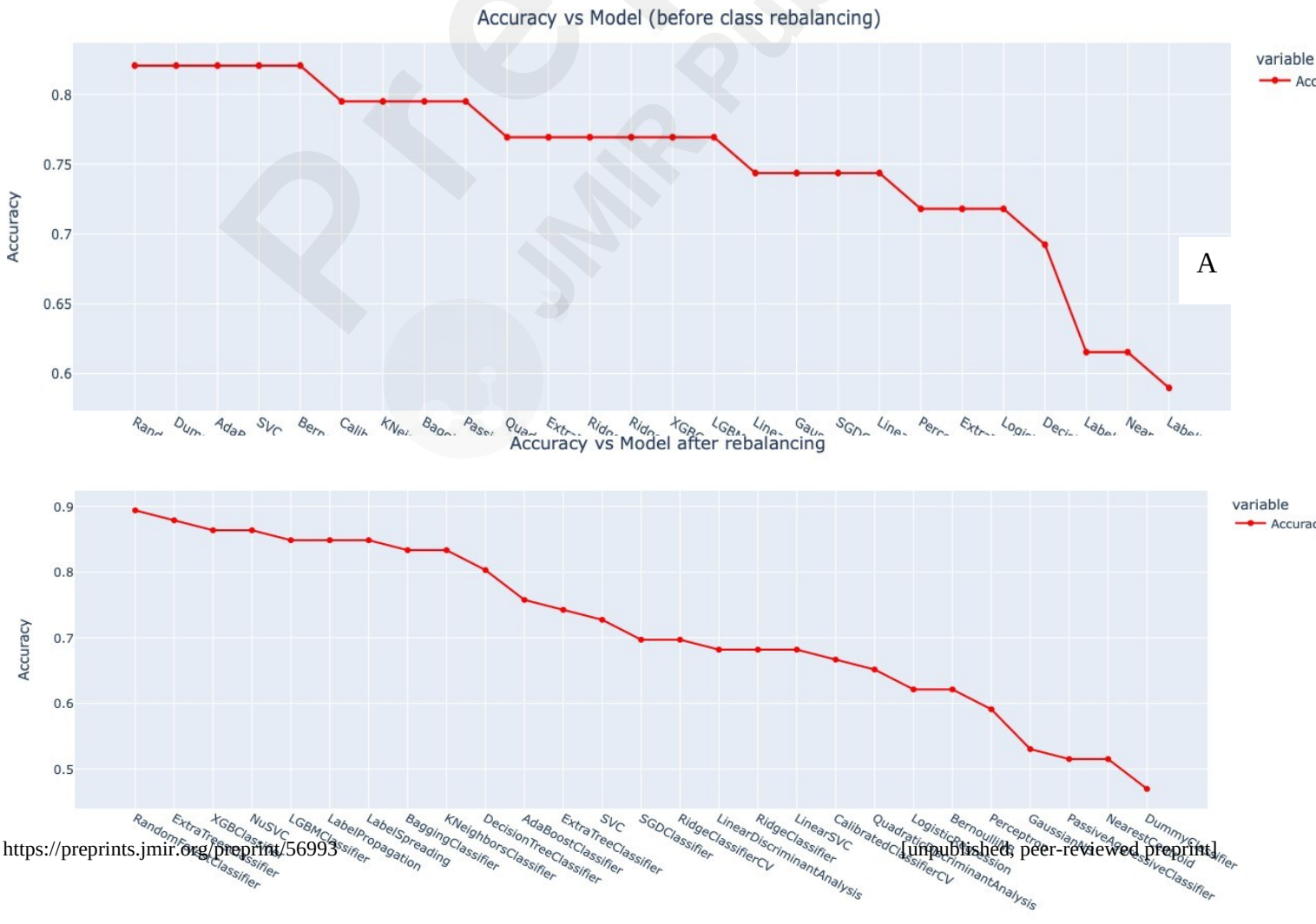


Figure 8: **Correlation matrix of independent variables with the outcome variable**



B

Figure 9: Accuracy scores of machine learning classifiers (A) before class rebalancing with SMOTE (B) after class rebalancing with SMOTE

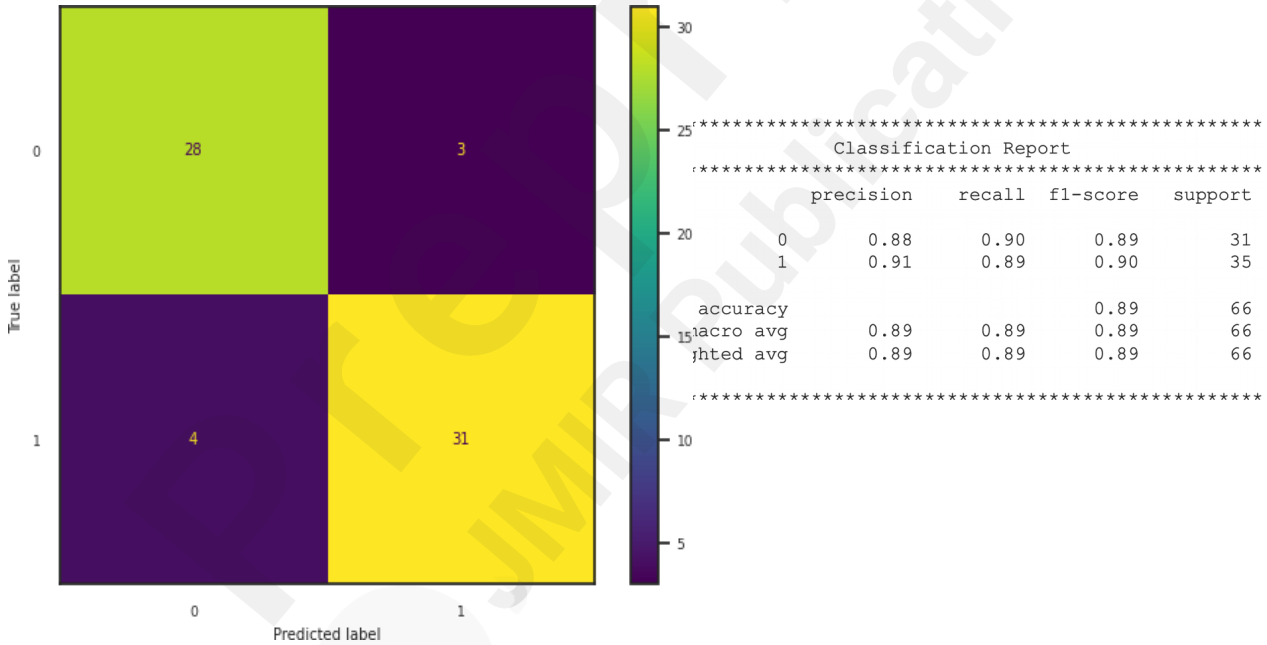


Figure 10: Random Forest confusion matrix indicating visual representation of the True vs Predicted Labels. True Positive (TP): The values which were positive and were predicted positive. That is, 31 cases of hyperglycemia were predicted correctly by the model. False Positive (FP): The values which were negative but falsely predicted as positive. In this case, only three cases were FP. False Negative (FN): The values which were positive but falsely predicted as negative. FN was four in this instance. True Negative (TN): The values which were negative and were predicted negative. Here, 28 cases were detected. In all, the weighted average of accuracy score = 0.89 and F1 Score = 0.89. *Precision* is a metric that quantifies the accuracy of a classifier by determining the number of correctly identified members of a class divided by all instances where the model predicted that specific class. In the context of hyperglycemia prediction, precision would be the count of accurate predictions of hyperglycemia divided by the total instances where the classifier predicted "hyperglycemia," regardless of correctness. *Recall*, on the other hand, measures the effectiveness of a classifier in correctly identifying members of a class by dividing the number of correctly identified instances by the total number of actual members in that class. In the hyperglycemia scenario, recall would represent the number of actual hyperglycemic individuals correctly identified

by the classifier. The *F1* score is a composite metric that combines both precision and recall into a single value. It provides a concise evaluation of a classifier's performance. A high *F1* score indicates that both precision and recall are high, while a low *F1* score suggests that one or both metrics are low. This metric is particularly useful for quickly assessing whether a classifier effectively identifies members of a class or if it resorts to shortcuts, such as indiscriminately classifying everything as a member of a larger class.

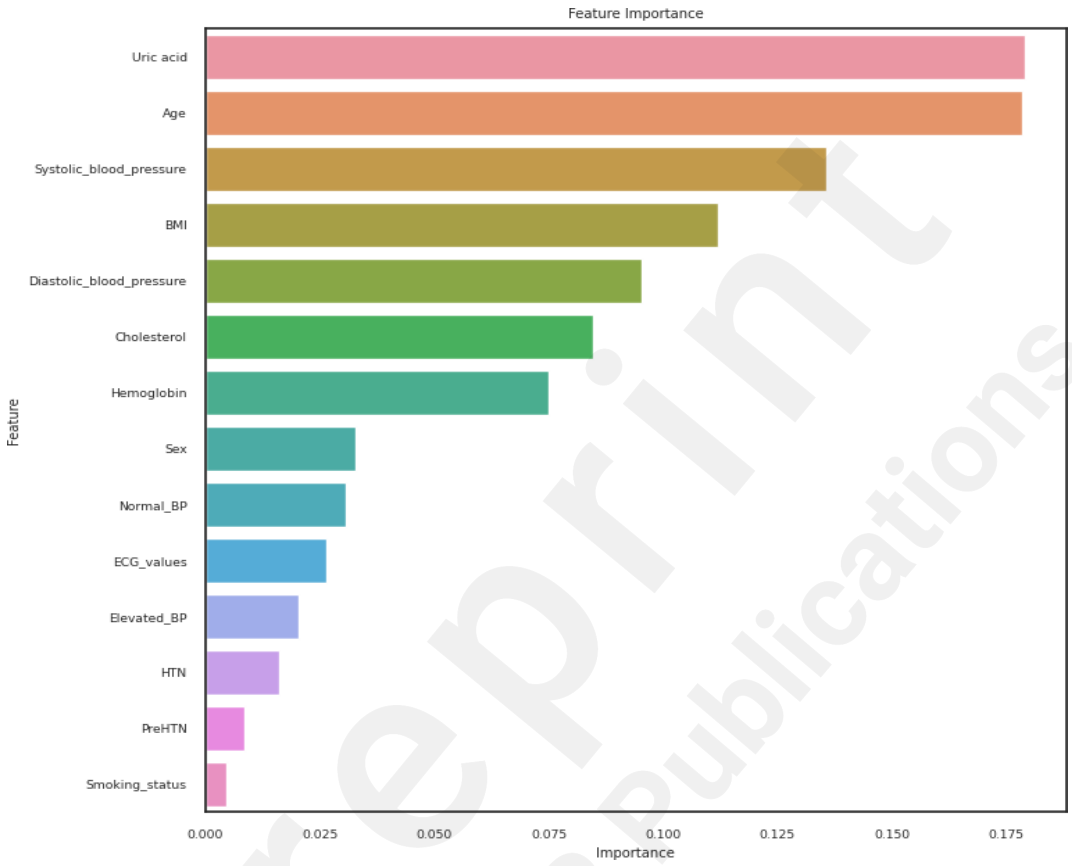


Figure 11: Blood glucose risk predictors

References

1. Bigna JJ, Noubiap JJ. The rising burden of non-communicable diseases in sub-Saharan Africa. Lancet Glob Health. 2019;7(10):e1295-e6.
2. Cross SH, Mehra MR, Bhatt DL, Nasir K, O'Donnell CJ, Califf RM, et al. Rural-Urban Differences in Cardiovascular Mortality in the US, 1999-2017. JAMA. 2020;323(18):1852-4.
3. Turecamo SE, Xu M, Dixon D, Powell-Wiley TM, Mumma MT, Joo J, et al. Association of Rurality With Risk of Heart Failure. JAMA Cardiology. 2023;8(3):231-9.
4. Khayat S, Dolatian M, Navidian A, Mahmoodi Z, Sharifi N, Kasaeian A. Lifestyles in suburban populations: A systematic review. Electron Physician. 2017;9(7):4791-800.
5. Kolié D, Van De Pas R, Codjia L, Zurn P. Increasing the availability of health workers in rural sub-Saharan Africa: a scoping review of rural pipeline programmes. Human Resources for Health. 2023;21(1):20.
6. Ngene NC, Khaliq OP, Moodley J. Inequality in health care services in urban and rural settings in South Africa. Afr J Reprod Health. 2023;27(5s):87-95.
7. Jane Ling MY, Ahmad N, Aizuddin AN. Risk perception of non-communicable diseases: A systematic review on its assessment and associated factors. PLoS One. 2023;18(6):e0286518.

8. Tohidinezhad F, Khorsand A, Zakavi SR, Rezvani R, Zarei-Ghanavati S, Abrishami M, et al. The burden and predisposing factors of non-communicable diseases in Mashhad University of Medical Sciences personnel: a prospective 15-year organizational cohort study protocol and baseline assessment. BMC Public Health. 2020;20(1):1637.
9. Alanazi R. Identification and Prediction of Chronic Diseases Using Machine Learning Approach. J Healthc Eng. 2022;2022:2826127.
10. Park DJ, Park MW, Lee H, Kim Y-J, Kim Y, Park YH. Development of machine learning model for diagnostic disease prediction based on laboratory tests. Scientific Reports. 2021;11(1):7567.
11. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making. 2019;19(1):281.
12. Wang M, Ge W, Apthorp D, Suominen H. Robust Feature Engineering for Parkinson Disease Diagnosis: New Machine Learning Techniques. JMIR Biomed Eng. 2020;5(1):e13611.
13. Sampa MB, Biswas T, Rahman MS, Aziz NHBA, Hossain MN, Aziz NAA. A Machine Learning Web App to Predict Diabetic Blood Glucose Based on a Basic Noninvasive Health Checkup, Sociodemographic Characteristics, and Dietary Information: Case Study. JMIR Diabetes. 2023;8:e49113.
14. Sampa MB, Hossain MN, Hoque MR, Islam R, Yokota F, Nishikitani M, et al. Blood Uric Acid Prediction With Machine Learning: Model Development and Performance Comparison. JMIR Med Inform. 2020;8(10):e18331.
15. Abd El-Hafeez T, Shams MY, Elshaier Y, Farghaly HM, Hassanien AE. Harnessing machine learning to find synergistic combinations for FDA-approved cancer drugs. Sci Rep. 2024;14(1):2428.
16. Hassan E, Abd El-Hafeez T, Shams MY. Optimizing classification of diseases through language model analysis of symptoms. Scientific Reports. 2024;14(1):1507.
17. Keohane EM, Smith L, Walenga JM. Rodak's Hematology - E-Book: Rodak's Hematology - E-Book: Elsevier Health Sciences; 2015.
18. Yousefi M, Najafi Saleh H, Yaseri M, Jalilzadeh M, Mohammadi AA. Association of consumption of excess hard water, body mass index and waist circumference with risk of hypertension in individuals living in hard and soft water areas. Environ Geochem Health. 2019;41(3):1213-21.
19. Tan JL, Thakur K. Systolic Hypertension. StatPearls. Treasure Island (FL)2023.
20. Whelton PK, Carey RM, Aronow WS, Casey DE, Jr., Collins KJ, Dennison Himmelfarb C, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Circulation. 2018;138(17):e426-e83.
21. Diagnosis and classification of diabetes mellitus. Diabetes Care. 2010;33 Suppl 1(Suppl 1):S62-9.
22. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12(null):2825–30.
23. N. V. Chawla KWB, L. O. Hall, W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artifical Intelligence Research. 2002;16:321-57.
24. Buyya R, Hernandez SM, Kovvur RMR, Sarma TH. Computational Intelligence and Data Analytics: Proceedings of ICCIDA 2022: Springer Nature Singapore; 2022.
25. Lathkar M. High-Performance Web Apps with FastAPI. Apress Berkeley, CA. 2023.
26. Katende D, Kasamba I, Sekitoleko I, Nakuya K, Kusilika C, Buyinza A, et al. Medium-to-long term sustainability of a health systems intervention to improve service readiness and quality of non-communicable disease (NCD) patient care and experience at primary care settings in Uganda. BMC

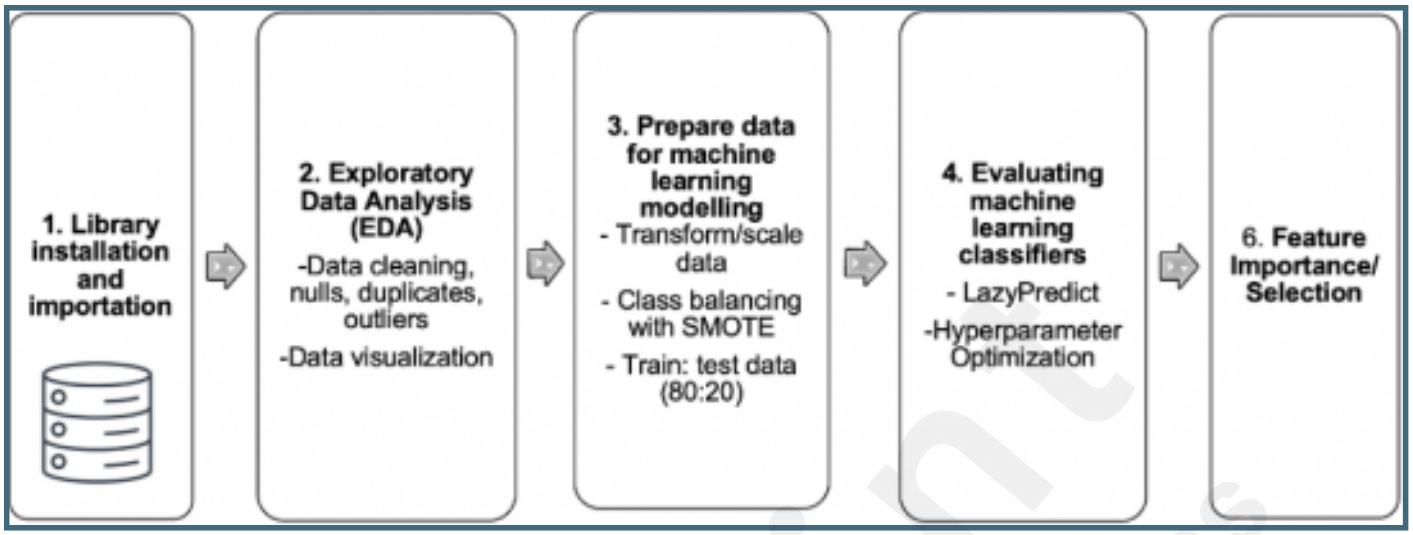
Health Serv Res. 2023;23(1):1022.
27. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019;6(2):94-8.
28. Abdel Hady DA, Abd El-Hafeez T. Predicting female pelvic tilt and lumbar angle using machine learning in case of urinary incontinence and sexual dysfunction. Scientific Reports. 2023;13(1):17940.
29. Eliwa EHI, El Koshiry AM, Abd El-Hafeez T, Farghaly HM. Utilizing convolutional neural networks to classify monkeypox skin lesions. Sci Rep. 2023;13(1):14495.
30. Farghaly HM, Shams MY, El-Hafeez TA. Hepatitis C Virus prediction based on machine learning framework: a real-world case study in Egypt. Knowl Inf Syst. 2023;65(6):2595–617.
31. Sharifi-Rad J, Rodrigues CF, Sharopov F, Docea AO, Can Karaca A, Sharifi-Rad M, et al. Diet, Lifestyle and Cardiovascular Diseases: Linking Pathophysiology to Cardioprotective Effects of Natural Bioactive Compounds. Int J Environ Res Public Health. 2020;17(7).
32. Liu R, Li D, Yang Y, Hu Y, Wu S, Tian Y. Systolic Blood Pressure Trajectories and the Progression of Arterial Stiffness in Chinese Adults. Int J Environ Res Public Health. 2022;19(16).
33. Singh JN, Nguyen T, Kerndt CC, Dhamoon AS. Physiology, Blood Pressure Age Related Changes. StatPearls. Treasure Island (FL): StatPearls Publishing
Copyright © 2023, StatPearls Publishing LLC.; 2023.
34. Song JJ, Ma Z, Wang J, Chen LX, Zhong JC. Gender Differences in Hypertension. J Cardiovasc Transl Res. 2020;13(1):47-54.
35. Wu J, Jiao B, Fan Y. Urbanization and systolic/diastolic blood pressure from a gender perspective: Separating longitudinal from cross-sectional association. Health Place. 2022;75:102778.
36. Midtbø H, Gerdts E. Sex disparities in blood pressure development: time for action. Eur J Prev Cardiol. 2022;29(1):178-9.
37. Fleg JL, Forman DE. Aging Changes in Cardiovascular Structure and Function. In: Waldstein SR, Kop WJ, Suarez EC, Lovallo WR, Katzel LI, editors. Handbook of Cardiovascular Behavioral Medicine. New York, NY: Springer New York; 2022. p. 127-62.
38. Fleg JL, Strait J. Age-associated changes in cardiovascular structure and function: a fertile milieu for future disease. Heart Fail Rev. 2012;17(4-5):545-54.
39. Hacker TA, McKiernan SH, Douglas PS, Wanagat J, Aiken JM. Age-related changes in cardiac structure and function in Fischer 344 × Brown Norway hybrid rats. American Journal of Physiology-Heart and Circulatory Physiology. 2006;290(1):H304-H11.
40. King H, Aubert RE, Herman WH. Global burden of diabetes, 1995-2025: prevalence, numerical estimates, and projections. Diabetes Care. 1998;21(9):1414-31.
41. Herman WH. The Global Burden of Diabetes: An Overview. In: Dagogo-Jack S, editor. Diabetes Mellitus in Developing Countries and Underserved Communities. Cham: Springer International Publishing; 2017. p. 1-5.
42. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet. 2023;402(10397):203-34.
43. Huang PL. A comprehensive definition for metabolic syndrome. Dis Model Mech. 2009;2(5-6):231-7.
44. Rafaqat S, Sharif S, Majeed M, Naz S, Manzoor F, Rafaqat S. Biomarkers of Metabolic Syndrome: Role in Pathogenesis and Pathophysiology Of Atrial Fibrillation. J Atr Fibrillation. 2021;14(2):20200495.
45. Srikanthan K, Feyh A, Visweshwar H, Shapiro JI, Sodhi K. Systematic Review of Metabolic Syndrome Biomarkers: A Panel for Early Detection, Management, and Risk Stratification in the West

Virginian Population. Int J Med Sci. 2016;13(1):25-38.
46. Madhusoodanan J. Searching for Better Biomarkers for Metabolic Syndrome. ACS Central Science. 2022;8(6):682-5.
47. Schonlau M, Zou RY. The random forest algorithm for statistical learning. The Stata Journal. 2020;20(1):3-29.
48. Ghaffar Nia N, Kaplanoglu E, Nasab A. Evaluation of artificial intelligence techniques in disease diagnosis and prediction. 2023;3(1).
49. Longo M, Bellastella G, Maiorino MI, Meier JJ, Esposito K, Giugliano D. Diabetes and Aging: From Treatment Goals to Pharmacologic Therapy. Frontiers in Endocrinology. 2019;10.
50. Yan Z, Cai M, Han X, Chen Q, Lu H. The Interaction Between Age and Risk Factors for Diabetes and Prediabetes: A Community-Based Cross-Sectional Study. Diabetes Metab Syndr Obes. 2023;16:85-93.
51. Nuredini G, Saunders A, Rajkumar C, Okorie M. Current status of white coat hypertension: where are we? Ther Adv Cardiovasc Dis. 2020;14:1753944720931637.
52. Franklin SS, Thijs L, Hansen TW, O'Brien E, Staessen JA. White-Coat Hypertension. Hypertension. 2013;62(6):982-7.
53. Luciano GL, Brennan MJ, Rothberg MB. Postprandial hypotension. Am J Med. 2010;123(3):281.e1-6.
54. Ali O. Genetics of type 2 diabetes. World J Diabetes. 2013;4(4):114-23.
55. Li C, Yang Y, Liu X, Li Z, Liu H, Tan Q. Glucose metabolism-related gene polymorphisms as the risk predictors of type 2 diabetes. Diabetology & Metabolic Syndrome. 2020;12(1):97.
56. Association WM. World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. JAMA. 2013;310(20):2191-4.

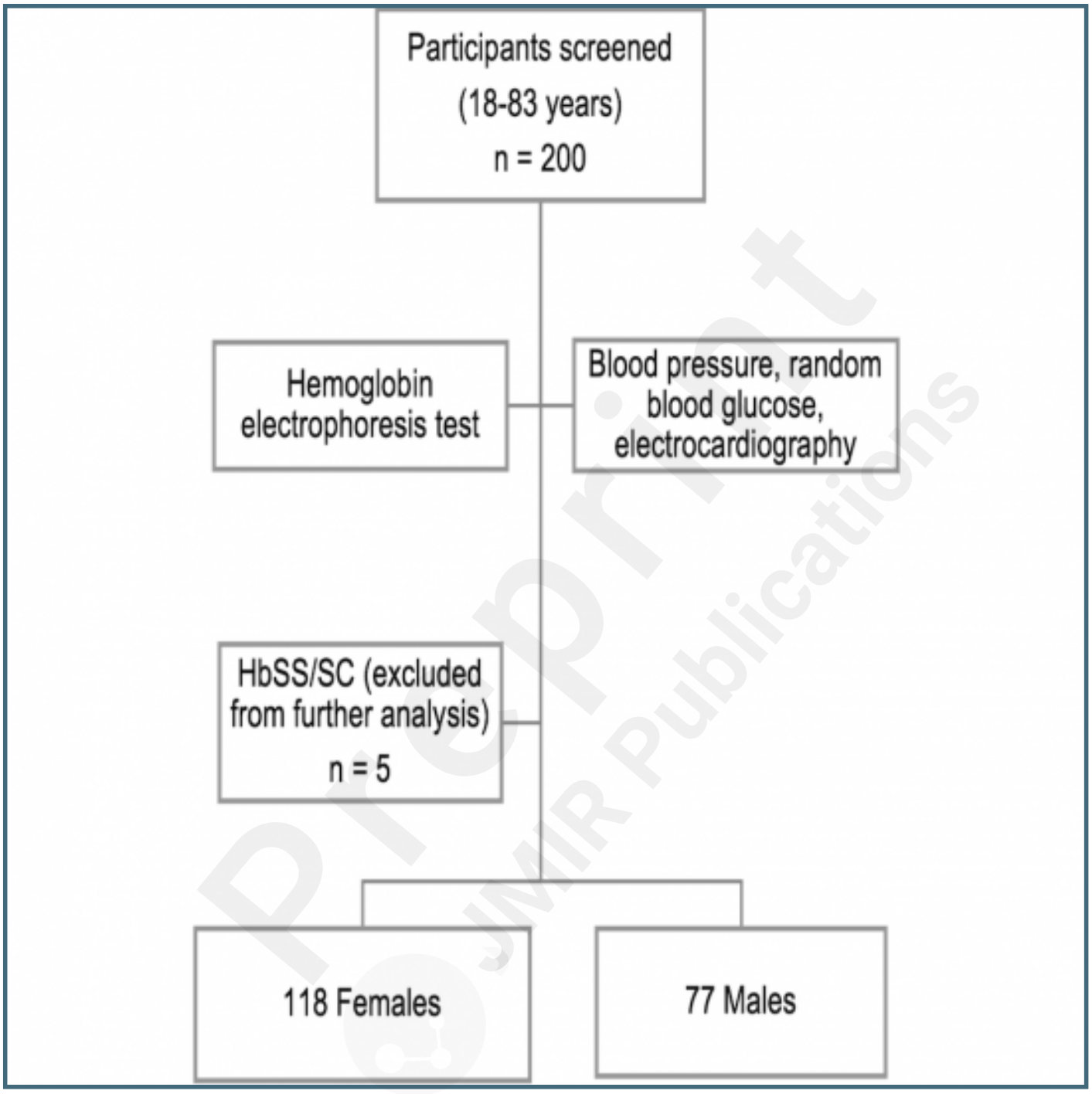
Supplementary Files

Figures

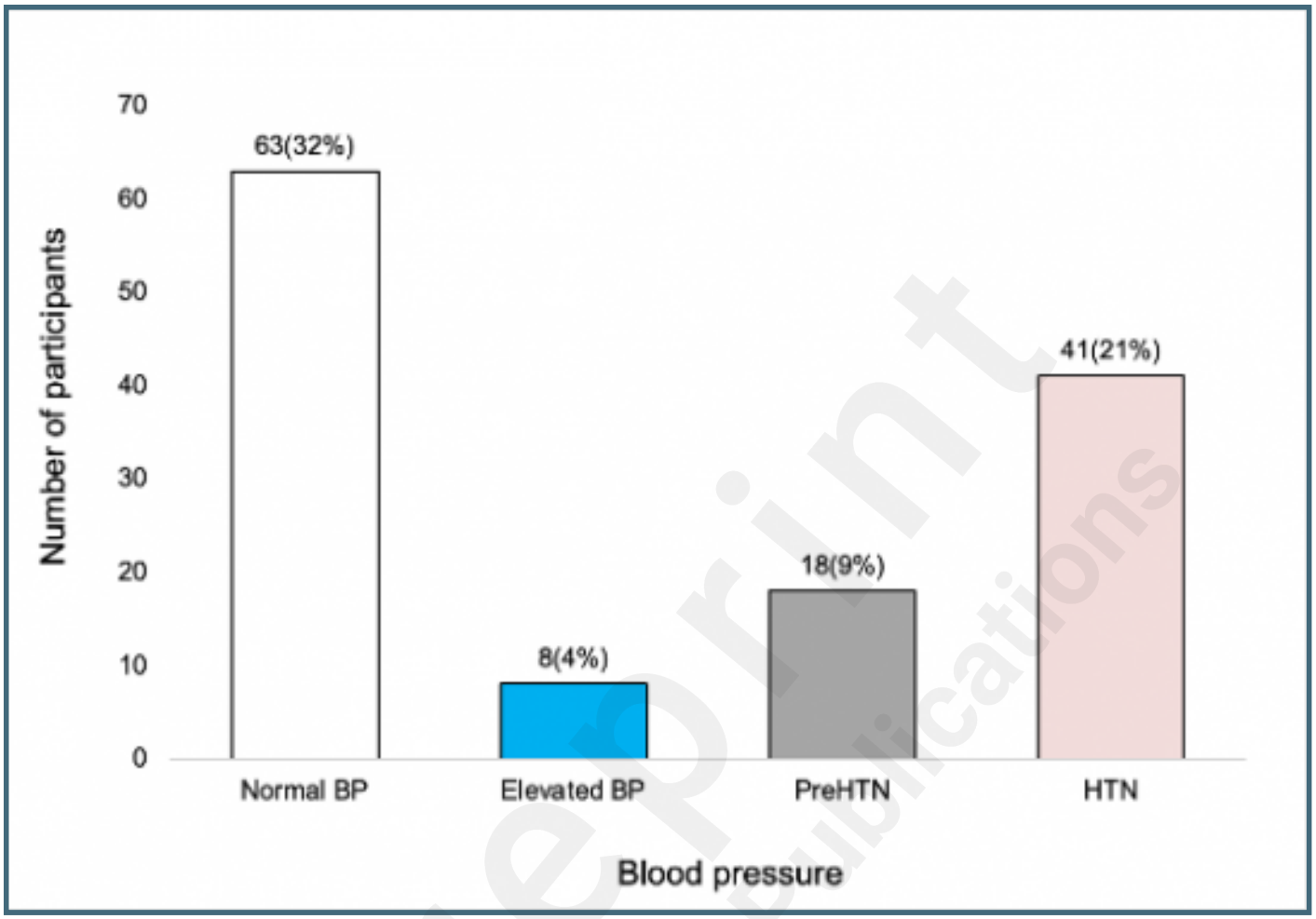
Pipeline for model development.



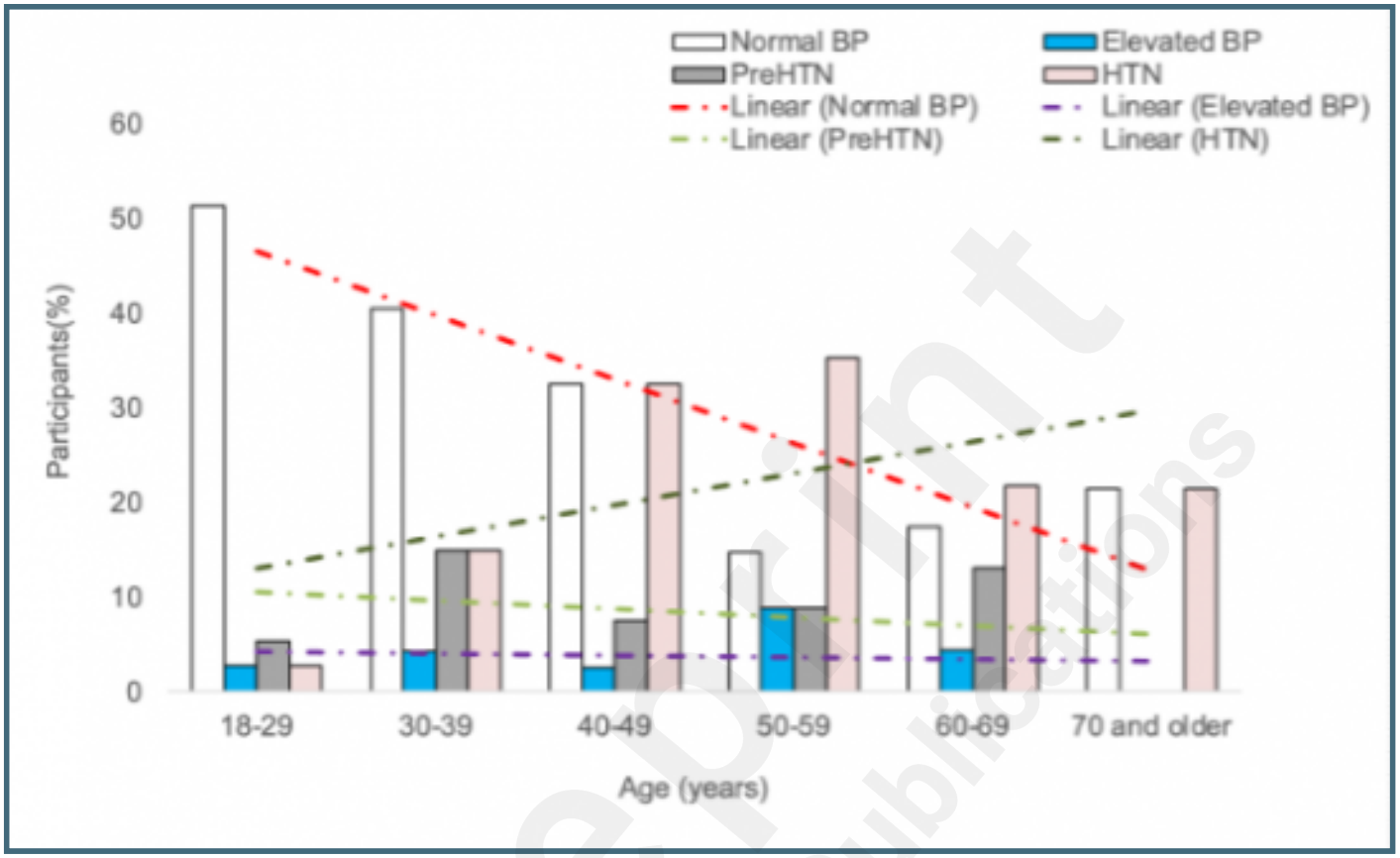
Participant recruitment and screening.



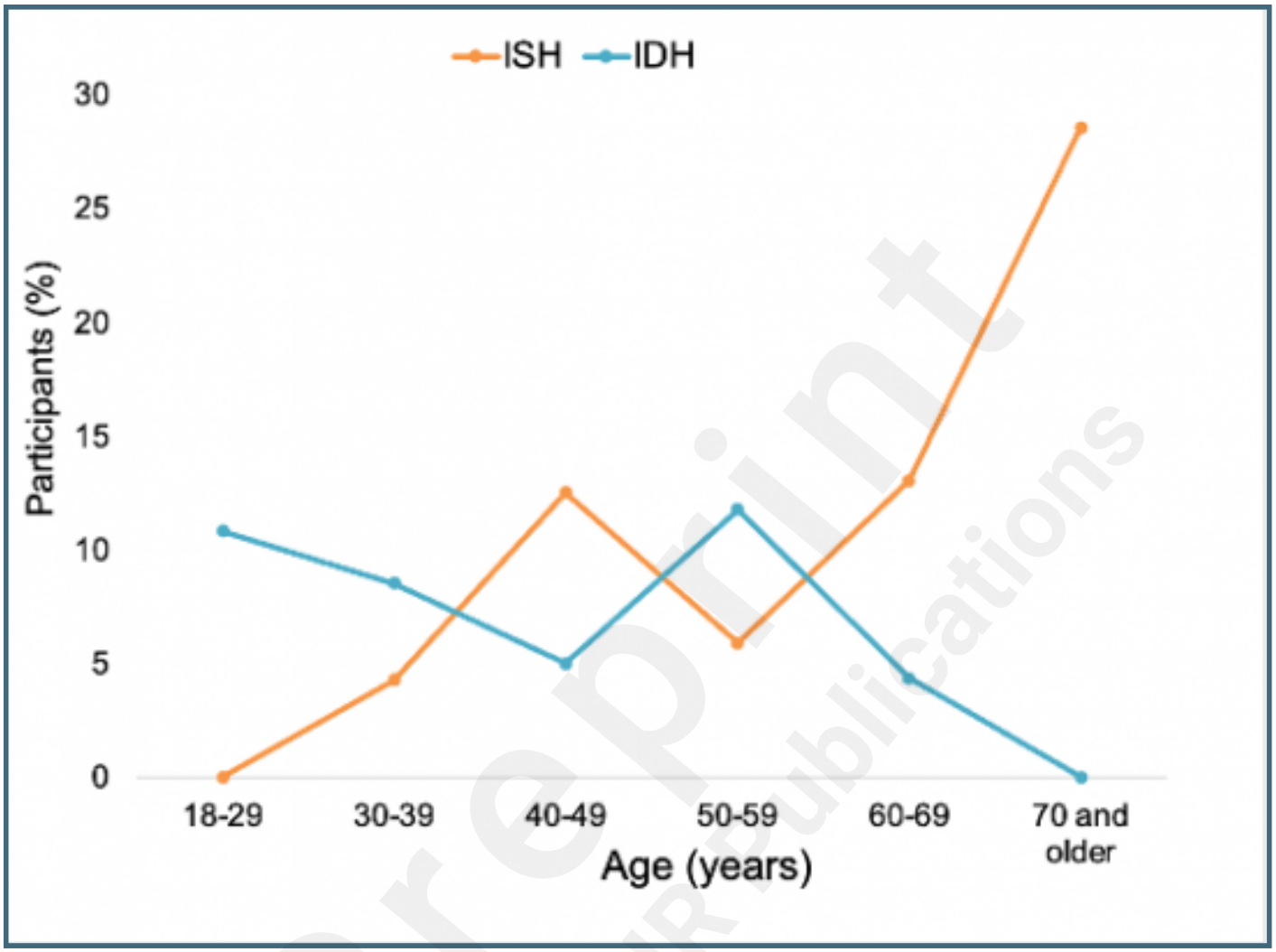
Blood pressure readings (BP = Blood Pressure; HTN = Hypertension).



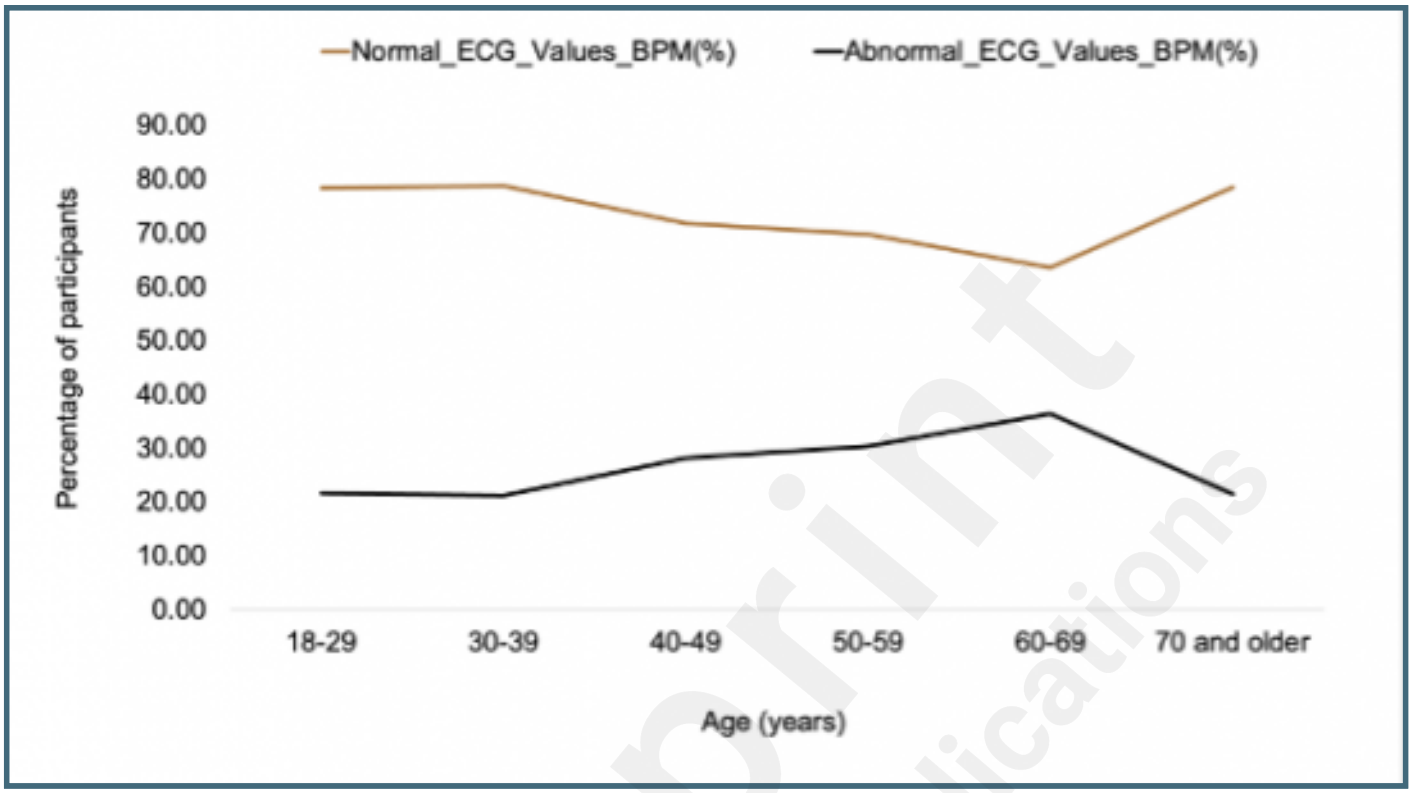
Age-based analysis of blood pressure. Percentage of participants with normal blood pressure (BP) reduced with increase in age (r = -0.88; P = .02). Prevalence of hypertension (HTN) increased with age (r = 0.53; P = .27), peaking between 50-59 years.



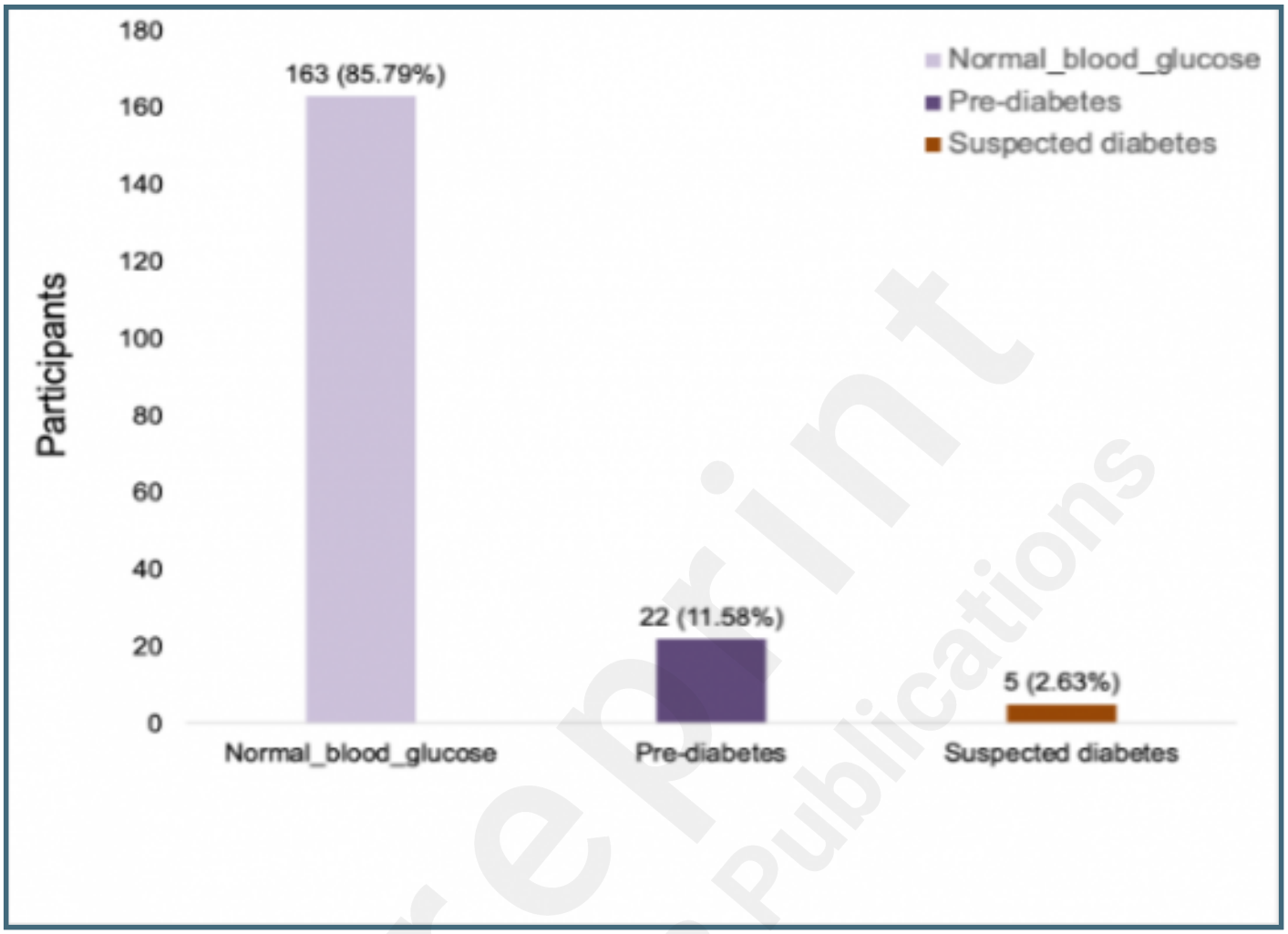
Age-based analysis of isolated systolic hypertension (ISH) and isolated diastolic hypertension (IDH). ISH increased with participants' age (r = 0.86; P = .03), unlike IDH (r = -0.71; P = .11).



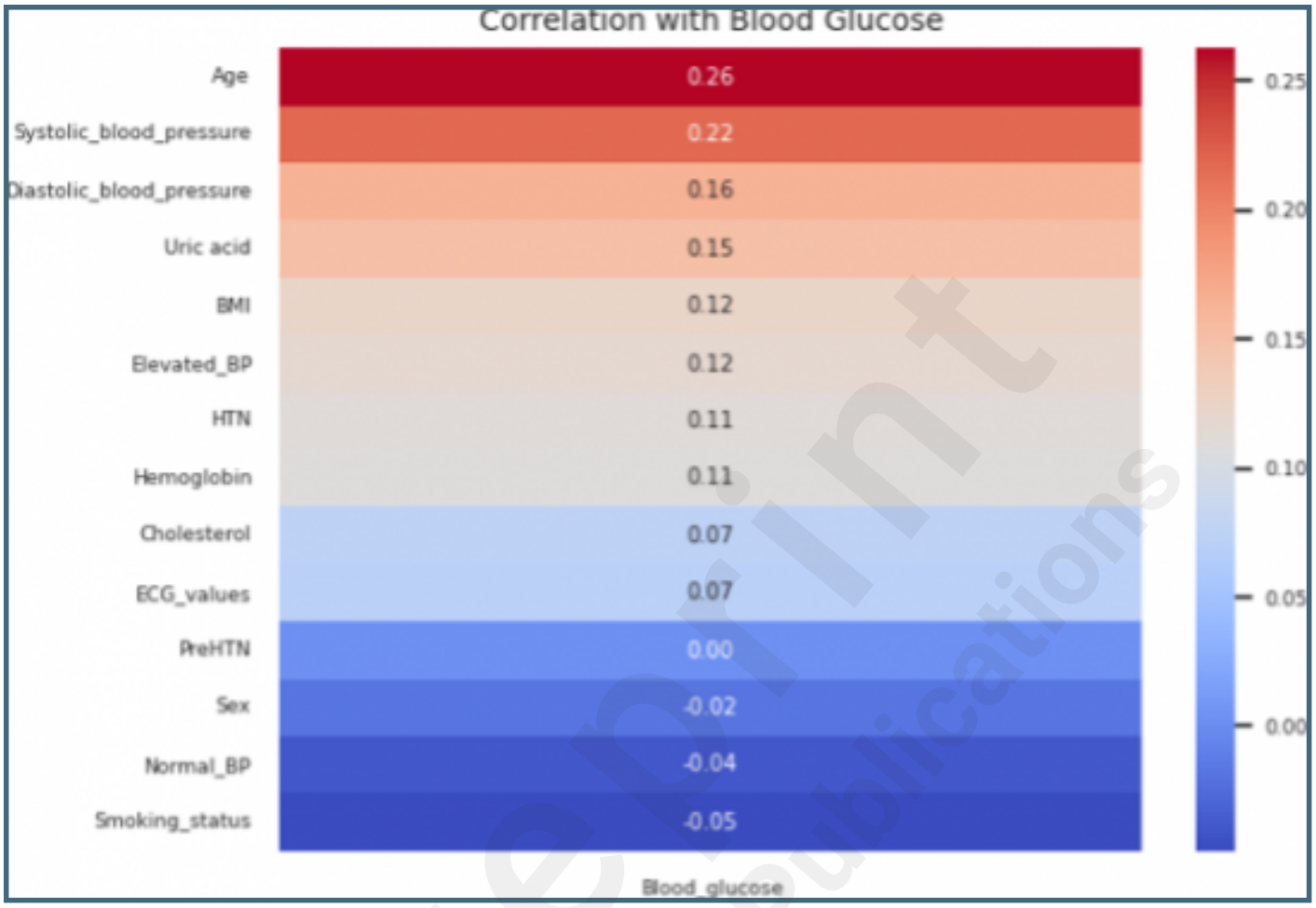
Age-based electrocardiogram (ECG) analysis. Age-dependent increase in the percentage of participants with abnormal ECG values peaking between 60-69 years.



Blood glucose levels in the cohort.

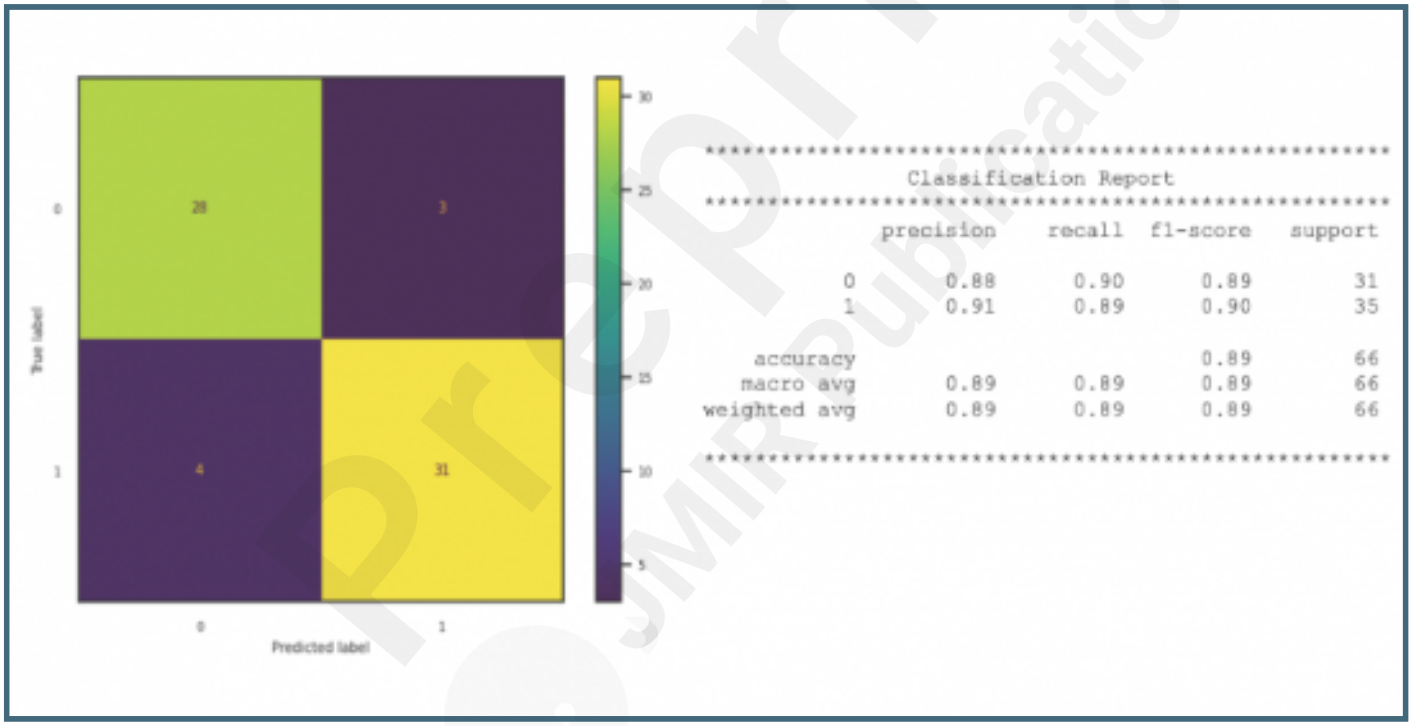


Correlation matrix of independent variables with the outcome variable.



Accuracy scores of machine learning classifiers (A) before class rebalancing with SMOTE (B) after class rebalancing with SMOTE.

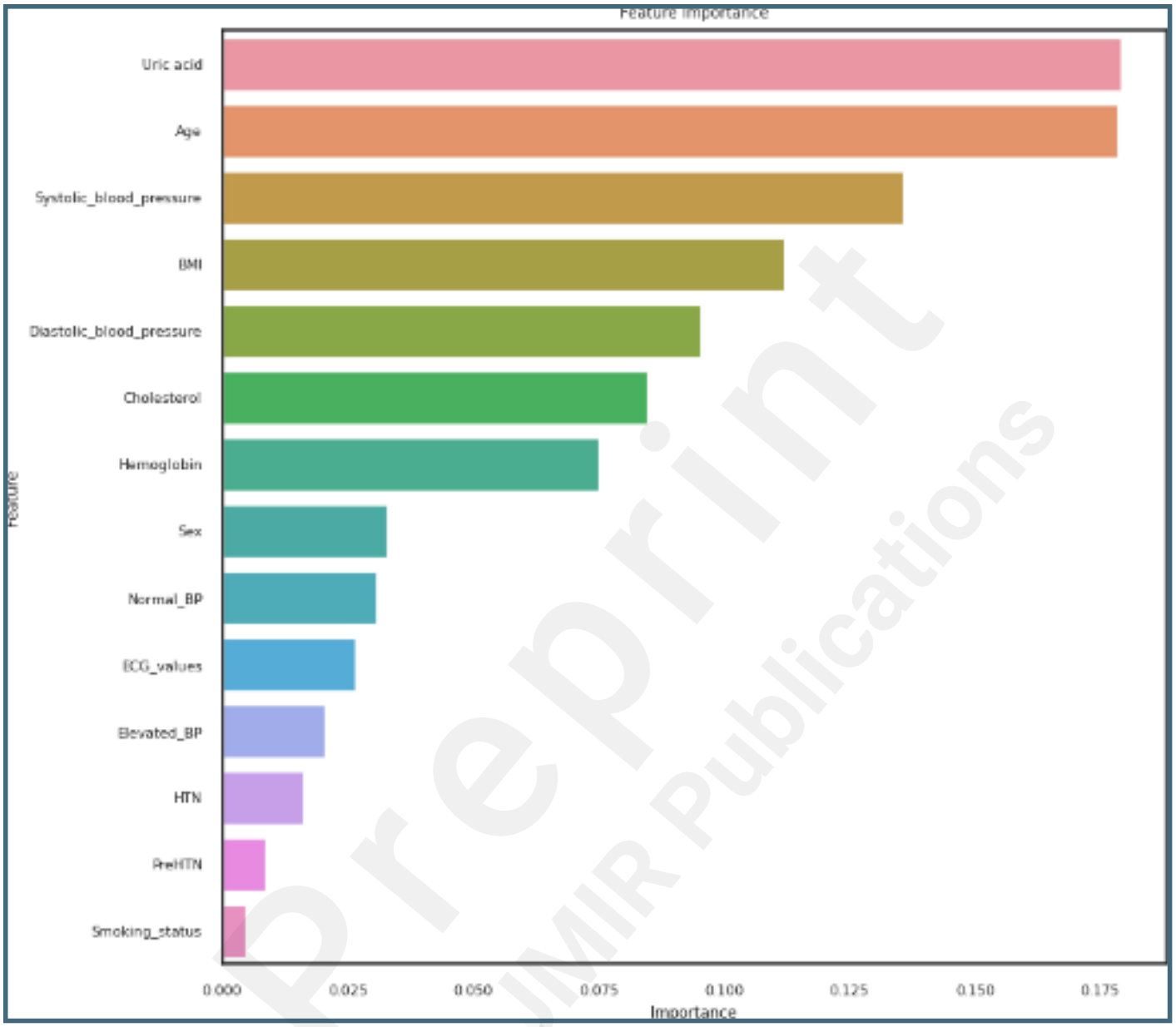
Random Forest confusion matrix indicating visual representation of the True vs Predicted Labels. True Positive (TP): The values which were positive and were predicted positive. That is, 31 cases of hyperglycemia were predicted correctly by the model. False Positive (FP): The values which were negative but falsely predicted as positive. In this case, only three cases were FP. False Negative (FN): The values which were positive but falsely predicted as negative. FN was four in this instance. True Negative (TN): The values which were negative and were predicted negative. Here, 28 cases were detected. In all, the weighted average of accuracy score = 0.89 and F1 Score = 0.89. Precision is a metric that quantifies the accuracy of a classifier by determining the number of correctly identified members of a class divided by all instances where the model predicted that specific class. In the context of hyperglycemia prediction, precision would be the count of accurate predictions of hyperglycemia divided by the total instances where the classifier predicted "hyperglycemia," regardless of correctness. Recall, on the other hand, measures the effectiveness of a classifier in correctly identifying members of a class by dividing the number of correctly identified instances by the total number of actual members in that class. In the hyperglycemia scenario, recall would represent the number of actual hyperglycemic individuals correctly identified by the classifier. The F1 score is a composite metric that combines both precision and recall into a single value. It provides a concise evaluation of a classifier's performance. A high F1 score indicates that both precision and recall are high, while a low F1 score suggests that one or both metrics are low. This metric is particularly useful for quickly assessing whether a classifier effectively identifies members of a class or if it resorts to shortcuts, such as indiscriminately classifying everything as a member of a larger class.



Classification Report

	precision	recall	f1-score	support
0	0.88	0.90	0.89	31
1	0.91	0.89	0.90	35
accuracy			0.89	66
macro avg	0.89	0.89	0.89	66
weighted avg	0.89	0.89	0.89	66

Blood glucose risk predictors.



Multimedia Appendixes

Supplementary methods and results.
URL: http://asset.jmir.pub/assets/3dbb37dcb6243d5967afe63c69a810bd.docx

Raw data file.
URL: http://asset.jmir.pub/assets/3d0c4e027f2273bdddb30d61dd049e3a.xls